Universidade do Minho

Universidade de Vigo

linguaMATICA

Volume 7, Número 1- Julho 2015

ISSN: 1647-0818

Volume 7, Número 1 – Julho 2015

LinguaMÁTICA

ISSN: 1647–0818

Conteúdo

Editorial

Iniciamos este ano de 2015 con dúas magníficas novas. Por unha banda, a incorporación da nosa revista á DBLP, a prestixiosa base de datos bibliográfica do ámbito das ciencias da computación mantida por Michael Ley na Universidade de Trier. Pola outra, a aceptación recibida da inclusión de Linguamática na importantísima base de datos de bibliografía científica de Scopus mantida por Elsevier, inclusión pendente da completa indización dos contidos da revista por parte da editora.

O noso obxectivo é seguir facendo desta revista o marco científico máis acaído para o envío de artigos de alta calidade no eido do procesamento das linguas ibéricas, e con ese obxectivo traballamos arreo para lograr tamén que a Linguamática sexa incluída nos índices bibliográficos principais a nivel local e internacional.

Porén, tamén nos parece relevante anovar e mellorar continuamente o deseño da revista e, neste senso, introducimos neste número unha nova portada de estilo actualizado e un formato renovado dos artigos da revista que inclúe útiles enlaces nas citas e nos URL na súa versión PDF.

Esperamos poder seguir ofrécendovos esta revista cada vez con mellores características e agradecemos sempre a vosa participación e a vosa vontade que divulgar os vosos traballos en Linguamática.

Xavier Gómez Guinovart
José João Almeida
Alberto Simões

Comissão Científica

Alberto Álvarez Lugrís,
Universidade de Vigo

Alberto Simões,
Universidade do Minho

Aline Villavicencio,
Universidade Federal do Rio Grande do Sul

Álvaro Iriarte Sanroman,
Universidade do Minho

Ana Frankenberg-Garcia,
University of Surrey

Anselmo Peñas,
Univers. Nac. de Educación a Distancia

Antón Santamarina,
Universidade de Santiago de Compostela

Antoni Oliver González,
Universitat Oberta de Catalunya,

Antonio Moreno Sandoval,
Universidad Autónoma de Madrid

António Teixeira,
Universidade de Aveiro

Arantza Díaz de Ilarraza,
Euskal Herriko Unibertsitatea

Arkaitz Zubiaga,
Dublin Institute of Technology

Belinda Maia,
Universidade do Porto

Carmen García Mateo,
Universidade de Vigo

Diana Santos,
Linguateca/Universidade de Oslo

Ferran Pla,
Universitat Politècnica de València

Gael Harry Dias,
Université de Caen Basse-Normandie

Gerardo Sierra,
Univers. Nacional Autónoma de México

German Rigau,
Euskal Herriko Unibertsitatea

Helena de Medeiros Caseli,
Universidade Federal de São Carlos

Horacio Saggion,
University of Sheffield

Hugo Gonçalo Oliveira,
Universidade de Coimbra

Iñaki Alegria,
Euskal Herriko Unibertsitatea

Irene Castellón Masalles,
Universitat de Barcelona

Joaquim Llisterri,
Universitat Autònoma de Barcelona

José João Almeida,
Universidade do Minho

José Paulo Leal,
Universidade do Porto

Joseba Abaitua,
Universidad de Deusto

Juan-Manuel Torres-Moreno,
Lab. Informatique d'Avignon - UAPV

Kepa Sarasola,
Euskal Herriko Unibertsitatea

Laura Plaza,
Complutense University of Madrid

Lluís Padró,
Universitat Politècnica de Catalunya

Marcos Garcia,
Universidade de Santiago de Compostela

María Inés Torres,
Euskal Herriko Unibertsitatea

Maria das Graças Volpe Nunes,
Universidade de São Paulo

Mercè Lorente Casafont,
Universitat Pompeu Fabra

Mikel Forcada,
Universitat d'Alacant

Pablo Gamallo Otero,
Universidade de Santiago de Compostela

Patrícia Cunha França,
Universidade do Minho

Rui Pedro Marques,
Universidade de Lisboa

Salvador Climent Roca,
Universitat Oberta de Catalunya

Susana Afonso Cavadas,
University of Sheffield

Tony Berber Sardinha,
Pontifícia Univ. Católica de São Paulo

Xavier Gómez Guinovart,
Universidade de Vigo

Artigos de Investigação

Geração de Linguagem Natural para Conversão de Dados em Texto Aplicação a um Assistente de Medicação para o Português

Trainable NLG for Data to Portuguese – With application to a Medication Assistant

José Casimiro Pereira
Instituto Politécnico de Tomar
Portugal
casimiro@ipt.pt

António Teixeira
Dep. Electrónica Telec. & Informática/IEETA
Universidade de Aveiro, Portugal
ajst@ua.pt

Resumo

Novos equipamentos como 'smartphones' ou 'tablets' têm revolucionado a interacção do ser humano com a tecnologia, proporcionando novos desafios e oportunidades. Estes novos dispositivos são multimodais por natureza. De entre as várias modalidades, são particularmente interessantes as relacionadas com a interacção por voz e texto. Para que estas formas de interação possam ser usadas entre sistemas e utilizadores humanos, é essencial a existência de módulos capazes de traduzir as informações internas das aplicações em frases ou textos, para visualização no ecrã ou para serem sintetizados de forma a serem ouvidos. É, também, essencial que estes módulos possam gerar frases e textos nas línguas nativas dos utilizadores; que o processo de desenvolvimento não implique grandes conhecimentos e recursos, incluindo tempo de desenvolvimento; e o resultado da geração apresente a variabilidade necessária.

O objectivo principal é o de propor, implementar e avaliar um método de conversão de Dados-para-português passível de ser desenvolvido com um mínimo de tempo e conhecimentos, mas sem comprometer a indispensável variabilidade e qualidade do que é gerado. O sistema apresentado, desenvolvido para um cenário de assistência à toma de medicamentos, destina-se a criar descrições, em linguagem natural, de informação sobre medicação a tomar. Motivados por resultados recentes, optou-se por uma abordagem baseada em tradução automática, com os modelos treinados num pequeno corpus paralelo.

Para isso, foi criado um novo corpus que, depois de validado, foi utilizado no desenvolvimento do sistema. Foram criadas duas variantes do sistema: uma orientada à tradução baseada em sintagmas e outra fazendo uso de informação sintáctica. Foram realizadas avaliações utilizando métricas automáticas – BLEU e Meteor – bem como avaliações por humanos. Os resultados do sistema orientado a sintagmas foram francamente superiores aos do seu concorrente, obtendo uma média por avaliador humano de 60% de frases consideradas inteligíveis, contra 46% do seu congénere, o que pode considerar-se um bom resultado tendo em conta a dimensão do corpus.

Palavras chave

Geração de linguagem natural, Dados-para-Texto, tradução automática, assistência à toma de medicação

Abstract

New equipments, such as smartphones and tablets, are changing human computer interaction. These devices present several challenges, especially due to their small screen and keyboard. In order to use text and voice in multimodal interaction, it is essential to deploy modules to translate the internal information of the applications into sentences or texts, in order to display it on screen or synthesize it. Also, these modules must generate phrases and texts in the user's native language; the development should not require considerable resources; and the outcome of the generation should achieve a good degree of variability.

Our main objective is to propose, implement and evaluate a method of data conversion to Portuguese which can be developed with a minimum of time and knowledge, but without compromising the necessary variability and quality of what is generated. The developed system, for a Medication Assistant, is intended to create descriptions, in natural language, of medication to be taken. Motivated by recent results, we opted for an approach based on machine translation, with models trained on a small parallel corpus.

For that, a new corpus was created. With it, two variants of the system were trained: phrase-based translation and syntax-based translation. The two variants were evaluated by automatic measurements – BLEU and Meteor – and by humans. The results showed that a phrase-based approach produced better results than a syntax-based one: human evaluators evaluated 60% of phrase-based responses as good, or very good, compared to only 46% of syntax-based responses. Considering the corpus size, we judge this value (60%) as good.

Keywords

Natural Language Generation, NLG, data2text, machine translation, medication assistant

1 Introdução

1.1 Motivação

O aumento do uso de dispositivos móveis, como *Smartphones*, *tablets* ou pequenos computadores, é, hoje, uma realidade indesmentível . Uma das principais dificuldades da sua utilização resulta das reduzidas dimensões do teclado e do ecrã. Estas características constituem ao mesmo tempo um desafio e uma oportunidade para o surgimento de novas tecnologias e interfaces.

Estes novos dispositivos são multimodais por natureza, uma vez que permitem várias formas de interacção: texto, imagens, voz, toque, vibração, etc..

De entre estas várias modalidades, são particularmente interessantes as relacionadas com a interacção por voz e texto. Para que tal seja possível, é essencial a existência de módulos capazes de traduzir as informações internas das aplicações em frases ou textos, para visualização no ecrã ou para serem sintetizadas de forma a serem ouvidas pelo utilizador.

Como requisitos adicionais, mas essenciais, temos:

(1) A necessidade desses módulos gerarem frases e textos com a suficiente variedade/variabilidade ao longo do tempo – uma das importantes características das frases produzidas pelos humanos – para que não se tornem aborrecidos e, em consequência, sejam considerados pouco naturais e utilizáveis;

(2) A possibilidade de criar módulos adequados para um número crescente de aplicações, sem serem necessários grandes conhecimentos de áreas não dominadas pelos *developers* de aplicações em geral(como conhecimentos aprofundados de Linguística), e sem ser necessário um grande investimento em termos de tempo de desenvolvimento;

(3) Utilização da língua portuguesa (quinta língua mais falada no mundo), abrindo portas à utilização deste tipo de aplicações na sua língua nativa a cerca de 240 milhões de pessoas (Alves, 2011), com expectativas de crescimento para perto de 335 milhões, em 2050 (Agência Lusa, 2010).

Muitos têm sido os esforços no sentido de dotar os computadores em geral, e estes dispositivos móveis em particular, da capacidade de "falarem" com os seres humanos (Jurafsky & Martin, 2009). Um dos primeiros esforços foi a criação de interacção através de frases e textos pré-definidos ou utilizando mensagens de voz pré-gravadas. Se, por um lado, a opção por modelos pré-definidos (vulgo *templates*) pode permitir a existência de sistemas simples de uma forma rápida, por outro, sem um investimento grande na criação de um número elevado de *templates*, teremos uma indesejável repetição das frases e textos produzidas pelo sistema. Esta repetição sistemática de uma mesma frase tipo resulta em sistemas percepcionados pelos utilizadores como pouco naturais, reduzindo a sua usabilidade e aceitação por parte destes.

Com o tempo foram propostas formas alternativas de geração de frases e texto, como os sistemas clássicos de Geração automática de Língua Natural (GLN) – em inglês, Natural Language Generation, ou NLG (Reiter & Dale, 1997, 2000). Os sistemas de GLN precisam de mapear alguma fonte de informação (como uma base de dados, por exemplo) em algum tipo de mensagem gerada automaticamente (Bateman & Zock, 2004). No entanto, esta tarefa está longe de ser considerada trivial, necessitando o seu desenvolvimento de muitos recursos (conhecimentos, corpora e tempo). A estes sistemas é requerido que decidam "como" dizer, depois de terem decidido "o que" dizer (Lemon, 2010; Bateman & Zock, 2004). Isto significa que os sistemas de GLN devem imitar os seres humanos, produzindo mensagens que são sintáctica e semanticamente corretas, além de serem, também, contextualmente adequadas.

Apesar de já existirem algumas experiências bem-sucedidas na criação de sistemas (Hunter et al., 2005; Konstantopoulos et al., 2008; McCauley et al., 2008), os recursos necessários para o desenvolvimento de sistemas de GLN clássicos genéricos continuam a ser escassos e o processo moroso, requerendo conhecimentos aprofundados de áreas como a Linguística e Processamento de Linguagem Natural. Para além disso, a sua adaptação a novos requisitos é, em geral, bastante difícil (Lemon, 2010). Constata-se também, através de uma análise da literatura nesta área, que a maioria dos sistemas e recursos necessários foi desenvolvida para a língua inglesa. Em resumo, este tipo de sistemas não se apresenta capaz de cumprir com os requisitos apresentados, sendo necessário explorar alternativas, em especial as que permitam obter sistemas para português e sem um grande investimento em termos de recursos.

Atendendo a que em muitas aplicações concretas para ambientes móveis a parte inicial do problema de geração – "o que" dizer – se encontra resolvido, apenas se torna necessário um sistema mais simples. Esta variante, designada habitualmente por sistemas de conversão de Dados-Para-

Texto (em inglês, Data2Text) (Reiter, 2007), utiliza como fonte para a geração um recurso não linguístico, frequentemente informação interna à aplicação, como dados de alguma fonte de dados.

Conjugando o atrás exposto, o nosso objectivo principal é o de propor, implementar e avaliar um método de conversão de Dados-Para-Português, passível de ser desenvolvido com um mínimo de tempo e conhecimentos, mas sem comprometer a indispensável variabilidade e qualidade do que é gerado (ex: frases).

Para que se possa atingir esse objectivo, com base numa análise de sistemas recentes Data2Text, será explorada a capacidade de sistemas baseados na utilização de aprendizagem automática de uma tradução entre a informação interna e português, tendo por base um corpus paralelo.

1.2 Cenário de aplicação escolhido

Tendo em conta o envelhecimento acentuado da população, o estudo centrou-se neste este grupo de utilizadores. Devido à sua idade, as suas capacidades motoras e cognitivas são mais reduzidas, pelo que a introdução deste tipo de tecnologias, com interfaces em língua natural, oral e escrita, tende a facilitar a sua vida diária. É, também, potenciadora da diminuição do isolamento, da exclusão e do aumento da capacidade de trabalho e autonomia (Teixeira et al., 2013b).

O cenário idealizado refere-se a uma situação onde uma pessoa está a tomar medicamentos. Neste contexto, o sistema deve interagir com o utilizador, usando, como meio de comunicação, a língua portuguesa, assistindo-o nas suas necessidades. Por exemplo, se o utilizador perguntar pelo próximo medicamento a tomar, deverá receber como resposta a informação pretendida, bem como informação complementar que o ajude na sua decisão.

O artigo encontra-se organizado da seguinte forma: uma primeira secção, com uma descrição da motivação e objectivos para a realização deste trabalho, seguida da descrição do cenário da aplicação. Na secção 2, é feita uma breve descrição de alguns exemplos de sistemas de Dados-para-Texto e formas de os avaliar. De seguida, é feita uma descrição da arquitectura geral do sistema implementado. A secção 4 apresenta informação sobre o corpus criado e os desenvolvimentos efectuados para o preparar para estas experiências, seguindo-se, na secção 5, a descrição dos passos efectuados no treino dos diversos sistemas. A secção 6 apresenta os resultados da avaliação e informação sobre as primeiras utilizações. O artigo termina com discussão e conclusões, analisando criticamente os resultados e apontando perspectivas de trabalhos futuros.

2 Trabalho Relacionado

2.1 Exemplos de sistemas de conversão de Dados-para-Texto

Nesta secção, são apresentados alguns exemplos de sistemas orientados a Dados-Para-Texto. Apesar de existirem diversos trabalhos na área da Geração de Língua Natural em português (Oliveira, 2012; Fonseca, 1993; Mendes, 2004; Ribeiro, 1995; Soares, 2001), apenas conseguimos identificar dois estudos relativos ao tema deste subtipo de GLN.

Pollen Forecast for Scotland – é um sistema que pretende traduzir, em texto, uma previsão para a concentração de pólen, nas diversas zonas da Escócia (Turner et al., 2006), de modo a que as pessoas sensíveis a níveis de pólen elevados possam precaver-se. Utiliza um corpus alinhado de 69 pares de frases, correspondentes a níveis de concentração e descrições, escritas por pessoas, referentes a essas concentrações. Este projecto surge como continuação do projecto Sumtime (Hunter et al., 2005), que efectua descrições textuais de previsões de meteorologia, em função dos dados meteorológicos fornecidos.

BabyTalk – Este sistema (Portet et al., 2009; Hunter et al., 2011) surgiu com o objectivo de apoiar os profissionais de saúde (enfermeiros e médicos) de uma Unidade de Cuidados Intensivos Neonatais. Efectivamente, estes profissionais, ao entrarem no seu turno, têm necessidade de assimilar uma grande quantidade de informação, em muito pouco tempo, sobre os bebés aí internados. Essa informação, normalmente, está distribuída por uma grande quantidade de dados sobre os bebés (análises de laboratório, dados dos equipamento de apoio à vida, dados sobre intervenções anteriores, etc.). O BabyTalk proporciona a esses profissionais resumos dos dados relevantes, tornando mais fácil e rápida a assimilação da informação prestada. Como corolário deste projecto, o BabyTalk pretende construir um módulo que permita às famílias terem, em tempo útil, resumos do estado de saúde dos seus bebés (Hunter et al., 2011).

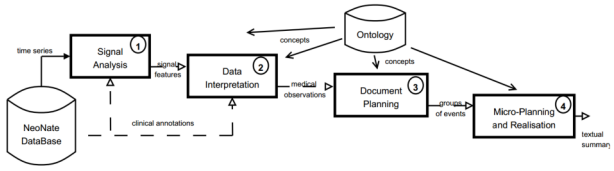


Figura 1: Arquitectura do BabyTalk (BT-45), retirado de (Portet et al., 2009).

SinNotas – É um dos poucos exemplos, de que temos conhecimento, de sistemas de Dados-Para-Texto, orientados para a língua portuguesa. Desenvolvido no Brasil, por Novais e Araújo (de Araújo et al., 2010; Novais et al., 2009), destina-se a dar apoio a uma aplicação de divulgação de notas de alunos, numa universidade. O SinNotas utiliza um corpus alinhado, onde a cada nota possível de um aluno se associa uma descrição para essa nota. Os autores defendem que, com este sistema, conseguiram que os alunos tivessem uma melhor percepção do entendimento dos professores sobre o seu desempenho.

Atributo	Descrição	Valores possíveis / Número de instâncias
provas_aval	Regular exams grades	nao_realizou(50), muito_abaixo(30), razoável(40), bom_mas_baixo(6), bom(84), muito_bom(19), excelente(12)
provas_turma	Same, as compared to the entire class	nulo(50), abaixo(100), acima(91)
progresso	Overall progress throughout the term	nulo(50), declinio(50), menor_meio(48), maior_meio(65), aumento(28)
sub_aval*	Substitutive exams grades	nulo(223), muito_abaixo(16), abaixo(2), acima(0)
sub_turma*	Same, as compared to the entire class	nulo(214), abaixo(11), acima(16)
eps_aval	Practical exercises grades	nao_realizou(56), muito_abaixo(2), razoável(5), bom_mas_baixo(2), bom(22), muito_bom(33), excelente(121)
dev_ep1	Whether exercises were compulsory	nulo(207), sim(34)
freq_aval	Attendance to the lectures	nulo(188), nenhuma(44), insuficiente(9)
corel_nota_falta*	Lower grades / attendance relation	nulo(215), sim(26)
mf_aval	Final term exams	muito_abaixo(81), razoável(41), bom_mas_baixo(5), bom(70), muito_bom(27), excelente(17)
mf_turma	Same, as compared to the entire class	nulo(58), abaixo(48), acima(135)
rec_aval	Recuperation exams grades	nulo(200), muito_abaixo(17), razoável(8), bom_mas_baixo(0), bom(16), muito_bom(0), excelente(0)
aband_rec*	Abandoned recuperation exams	nulo(235), sim(6)
rec_turma	Same, as compared to the entire class	nulo(204), abaixo(16), media(2), acima(19)

Tabela 1: Mensagens e possíveis valores do SINotas (extraído de (Novais et al., 2009)).

PortNLG – É um recente exemplo de um sistema, visando o português como língua de trabalho, desenvolvido por Silva Junior, Paraboni e Novais (Silva Junior et al., 2013). Consiste numa biblioteca JAVA concretizando um realizador superficial. Destina-se a gerar frases em português, tendo como entrada uma especificação abstracta da frase a ser construída.

Mountain – O sistema Mountain foi desenvolvido por Langner(Langner & Black, 2009; Langner, 2010), como parte da sua tese de doutoramento. O Mountain utiliza, também, um corpus alinhado, e foi, dos sistemas analisados, o primeiro a utilizar a ferramenta MOSES (Koehn et al., 2007; Koehn, 2014) como forma de gerar os textos a serem apresentados aos seus utilizadores. A sua 'linguagem de entrada' corresponde a uma sequência de códigos que representam a disponibilidade de um 'court' de ténis. A 'linguagem de saída' corresponde à 'tradução' desse código em inglês.

000000 d5 d3	friday evening is completely closed
100000 d2 t2	the ony time available is noon
111111 d4 t1	the court is open all morning
111111 d1 t3	you can reserve a court anytime on monday evening
100011 d5 t3	six, ten or eleven
010011 d3 t2	you an reserve a court at 1pm, 4pm and 5pm on wednesday
011001 d4 t3	any time but 6, 9 and 10
111011 d7 d2	afternoon except the 3pm block
111100 d1 t2	you can reserve a court is free anytime from noon until 3
110111 d6 t3	saturday evening, ooh, that

Tabela 2: Exemplo do corpus do Mountain (retirado de (Langner, 2010)).

2.2 Avaliação da Geração

Para a utilização efectiva do sistema, é necessário efectuar testes para garantir a sua qualidade. Contudo, a avaliação de sistemas de GLN ainda não é consensual (Hastie & Belz, 2014) e, ao longo dos últimos anos, têm surgido diversas propostas que se podem dividir em dois grandes grupos de avaliação: avaliação feita por seres humanos e avaliações automáticas, efectuadas por computador. Estudos sugerem que as avaliações efectuadas por seres humanos são geralmente melhores do que as automáticas, quando o objecto de estudo são textos de apoio à realização de tarefas, apesar de, em alguns casos particulares, tal possa não se verificar (Law et al., 2005). Apesar desta realidade, as avaliações automáticas têm vindo a ser cada vez mais utilizadas, em especial devido ao enorme custo, em termos de tempo e recursos económicos, que a avaliação por seres humanos acarreta.

Consideram-se, então, para este tipo de sistemas, essencialmente 3 tipos de avaliações: avaliação orientada à tarefa, avaliação por humanos e métricas automáticas. A avaliação orientada à tarefa consiste em produzir os textos e, depois, entregá-los às pessoas que os vão utilizar. O objectivo é avaliar o quanto esses textos ajudam as pessoas a efectuar as suas tarefas. Estas avaliações consomem muito tempo, são bastante dispendiosos e difíceis de concretizar, especialmente quando envolvem pessoas muito qualificadas (Portet et al., 2009). A avaliação por humanos é efectuada fornecendo o texto gerado a uma, ou mais pessoas, e solicitando a

sua avaliação sobre a utilidade e correcção desse texto. As métricas automáticas foram desenvolvidas para substituir as avaliações envolvendo humanos, devido às restrições que esse tipo de avaliação encerra. Este tipo de métricas efectua a comparação entre o texto produzido pelo sistema e texto escrito por humanos, tendo por base a mesma fonte inicial.

BLEU – É o acrónimo para BiLingual Evaluation Understudy (Papineni et al., 2002). O BLEU é um algoritmo que avalia a aproximação entre um texto gerado automaticamente e um texto, previamente obtido, gerado por um ser humano. Quanto mais próximos tiverem, maior qualidade terá o texto em avaliação. Esta avaliação é efectuada sobre os elementos individuais do texto gerado (frases ou partes de frases), comparando-os com um texto de referência, com boa qualidade. O índice desta avaliação é depois extrapolado para todo o texto. Factores como a inteligibilidade ou questões gramaticais não são tidos em consideração nesta métrica. O BLEU é expresso através de um número entre 0 e 1. Quanto maior o valor, maior a similitude com o texto de comparação. Devido à forma como o teste é realizado – comparação entre 'ngrams' –, a avaliação produz valores aceitáveis, quando o texto em avaliação é confrontado com todo o texto de referência e produz valores maus, quando confrontado com simples frases individuais.

Meteor – É o acrónimo de Metric for Evaluation of Translation with Explicit ORdering (Denkowski & Lavie, 2014, 2011). Semelhantemente ao BLEU, o Meteor aplica sobre cada frase gerada o algoritmo de avaliação. Este algoritmo cria um alinhamento entre os constituintes (palavras) da frase em teste e uma frase de referência. Por alinhamento, entende-se uma correspondência direta entre dois *unigrams*, um da frase em análise e outro da frase de referência. A correspondência pode ocorrer ao nível do reconhecimento exato da palavra, se forem sinónimas ou se derivarem de uma mesma palavra. A correspondência pode ocorrer, também, se a frase em avaliação for paráfrase de outra considerada válida.

O índice desta métrica é obtido pelo cálculo da média harmónica entre o número de alinhamentos considerados corretos e o número de alinhamentos possíveis, sendo que este segundo valor tem maior peso que o primeiro. Desenvolvida para minorar alguns dos problemas evidenciados pelo BLEU, esta métrica foca-se nas frases do corpus de forma individual, enquanto o BLEU avalia essencialmente o corpus como um todo.

3 Arquitectura Geral do Sistema

A arquitectura geral do sistema aqui descrito é apresentada na figura seguinte (Figura 2). A sua missão consiste em gerar mensagens em língua natural. É um *componente* de um sistema mais vasto, que se encontra em desenvolvimento.

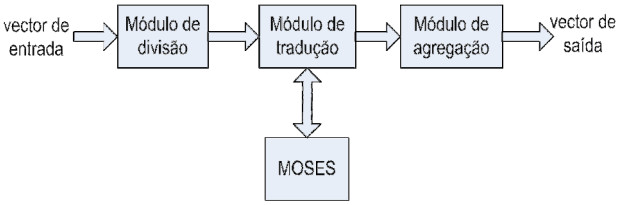


Figura 2: Arquitectura proposta

A parte central deste sistema, e objecto deste artigo, é um módulo – TRADUÇÃO – capaz de criar uma frase, em resposta a um vector com dados, fornecido como entrada. Se o vector fornecido como entrada for demasiado grande, será inicialmente dividido em diversos vectores. Posteriormente, um módulo de agregação irá juntar as diversas frases geradas, produzindo um texto coerente. Para se alcançar este objectivo são necessários três componentes.

O módulo base de dados (BD) é o componente responsável por armazenar todos os tipos de dados, desde as características dos utilizadores, características dos medicamentos, receitas médicas, etc..

O módulo MOSES é responsável pela tradução das frases, enviadas pelo módulo de TRADUÇÃO, para português. Para efectuar este serviço, o Moses precisa de efectuar o seu treino com um corpus, constituído por duas linguagens, perfeitamente alinhadas. A cada frase na linguagem de 'entrada' deve corresponder uma frase na linguagem de 'saída', respeitando o ordenamento dos ficheiros. A linguagem de 'entrada' será constituída por valores correspondentes aos fornecidos pelo módulo de BD. Na linguagem de 'saída' estão as expressões, em português, que se deseja que o Moses seja capaz de gerar.

O módulo de TRADUÇÃO é o módulo principal deste sistema. É responsável por receber os pedidos dos utilizadores e interagir com os dados armazenados na BD, guardando-os ou solicitando-os. É, também, sua responsabilidade enviar mensagens escritas na linguagem de 'entrada' para o módulo MOSES e receber a resposta na linguagem de 'saída'. Por último, compete-lhe processar as respostas e apresentá-las ao utilizador.

3.1 Dois tipos de tradução

O sistema MOSES suporta dois tipos de tradução muito diferentes, conhecidos pelas designações inglesas de *phrase-based* e *tree-based*. Adoptaremos neste artigo as designações de tradução baseada em *sintagmas* e tradução usando *sintaxe*.

Tradução baseada em sintagmas – As denominadas tabelas de tradução são a principal fonte de conhecimento para o *decoder*, que consulta estas tabelas para descobrir como traduzir uma entrada numa linguagem para uma outra linguagem, a de saída.

Estas tabelas de tradução não contêm apenas entradas correspondentes a uma palavra isolada, mas, em geral, as entradas são constituídas por múltiplas palavras. Deste facto deriva a designação de "baseada em sintagmas (frase)". No entanto, neste contexto 'frase' apenas significa uma sequência arbitrária de palavras.

Um exemplo possível de uma entrada na tabela, de acordo com o nosso cenário, seria:

```
Forma2 Medicamento1 ||| comprimidos de
MEDINX ||| 0.8 ||| |||
```

O processo de tradução consiste, como ilustrado na Figura 3, em dividir o vector de entrada em blocos para os quais exista uma tradução e, depois, combinar a saída de cada um desses blocos, com possível reordenação da posição.

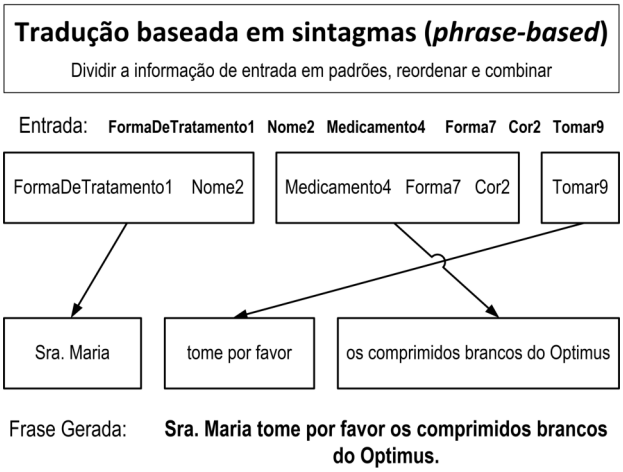


Figura 3: Esquema do funcionamento da tradução baseada em sintagmas

Tradução usando informação sintáctica – Estes modelos são também conhecidos como sistemas hierárquicos baseados em sintagmas e utilizam uma gramática do tipo SCFG (Synchronous Context-Free Grammar). Por esse motivo, quando a 'linguagem de saída' do corpus é processada, cada termo é classificado como 'não-terminal', associando a estes uma etiqueta que representa o seu conteúdo (Nome, Determinante, etc.). Por exemplo, nestes modelos sintácticos, os não-terminais são anotados com etiquetas com informação linguística, como verbo ('VERB') e nome ('NOUN'):

```
NOUN --> tipo1 ||| Comprimidos
VERB --> tomar2 ||| toma
```

Estas etiquetas são obtidas através da utilização de um *parser*, sendo depois utilizadas para efectuar o processo de alinhamento, entre as frases das linguagens de entrada e de saída.

Numa das variantes, que é a que interessa para o presente trabalho, as frases da linguagem de entrada permanecem inalteradas e apenas as de saída sofrem esta transformação. Esta variante é usualmente denominada de frase-para-árvore (string-to-tree, em inglês). As várias permutações – árvore-para-frase e árvore-para-árvore – são também possíveis.

Informação detalhada sobre estes dois tipos de sistemas, assim como sobre a forma de os construir, encontra-se disponível em dois tutoriais mantidos pelos responsáveis pelo sistema MOSES (MOSES, 2014b,c)

Treinados os sistemas, ficam disponíveis os modelos (de linguagem, de tradução, de reordenação) e parâmetros de configuração para a invocação do sistema para processamento de frases na linguagem de entrada. Embora para os dois tipos de sistemas atrás referidos exista um *decoder* MOSES específico, o processo de tradução é conceptualmente similar.

4 Corpus

4.1 Estrutura do corpus

Como referido anteriormente, foi utilizado um corpus constituído por duas linguagens, perfeitamente alinhadas. Por simplicidade, as duas linguagens utilizadas foram designadas por 'linguagem de entrada' e 'linguagem de saída'.

A 'linguagem de entrada' reflecte os dados que são obtidos por consulta, na base de dados, tendo sido seleccionados 9 tipos de dados: Nome e Apelido do utilizador, Mensagem de cortesia, Nome do medicamento a tomar, Tipo do medicamento, Forma de tomar o medicamento, Cor do medicamento, Dose a tomar e Frequência da toma. A estes termos foram concatenados os valores das chaves primárias, correspondentes aos registos seleccionados da base de dados. Desta forma, cada

frase reflecte os dados do utilizador e o que ele deve fazer.

Correspondentemente, na 'linguagem de saída' surge uma frase que exprime o mesmo tipo de informação da 'linguagem de entrada', mas em português. A Tabela 3 apresenta um exemplo destas duas linguagens.

4.2 Obtenção do corpus

O corpus utilizado nesta experiência foi obtido através do seguinte processo:

1. Foi solicitado a um conjunto de 15 voluntários, compreendidos entre os 18 e os 55 anos, que preenchessem um formulário para a introdução de frases, referentes à 'acção de informar uma pessoa sobre os medicamentos que deveria tomar'. Obtiveram-se 126 frases;
2. Aplicando uma estratégia semelhante à descrita em (Langner, 2010), o corpus original foi expandido para um total de 643 frases;
3. Esta expansão foi executada, porque se constatou que cada uma das frases, aplicadas a uma pessoa e medicamento concretos poderia, também, ser aplicada a outras pessoas ou medicamentos, desde que devidamente ajustadas ao novo contexto. Assim, foram executados os seguintes passos:
   (a) Mantendo a sequência original de introdução das frases, a cada frase foram acrescentados, manualmente, *tokens* que identificaram os termos que poderiam ser substituídos. A Tabela 4 apresenta um exemplo de duas frases recolhidas no corpus original e correspondente adaptação com os *tokens*. Na Tabela 5, são apresentados os diversos *tokens* utilizados.
   (b) Após esta fase, cada uma das 126 frases originais do corpus foi replicada entre 3 e 7 vezes, sendo que a geração do número de replicações foi efectuada de forma aleatória. Durante esta replicação, não foi alterada a sequência original do corpus.
   (c) Para permitir a substituição dos *tokens* por nomes de pessoas, medicamentos, etc. foi inicialmente criada uma pequena base de dados com 80 nomes próprios, 33 apelidos e 28 medicamentos. Para cada medicamento, foram identificados o seu nome, o tipo (comprimido, gotas, etc.) e a cor, quando aplicável. Foram, igualmente, identificados 11 períodos possíveis de 'toma' de medicamentos e 6 quantidades diferentes de medicamentos, para cada 'toma'.
   (d) Estas novas 643 frases correspondem à linguagem final do corpus. Ao mesmo tempo, e pela mesma sequência em que se substituía os *tokens* por valores concretos, foi criada uma nova lista de 643 frases, que correspondem à linguagem inicial. O objectivo foi gerar dois ficheiros, cujo conteúdo estivesse perfeitamente alinhado.
   (e) Para cada *token*, em cada frase, foi atribuído um valor escolhido aleatoriamente dentro da base de dados referida. Quando os *tokens* estavam relacionados (MEDICAMENTO, TIPO e COR, por ex.), a escolha de um medicamento implicou a escolha automática para os outros *tokens*, para garantir a integridade da informação.
   (f) A fase final consistiu numa análise, por uma pessoa, das frases obtidas através deste processo de expansão. Foram apenas corrigidos erros gramaticais.

4.3 Preparação do corpus para treino e teste

Usando a técnica *10-fold cross-validation*, descrita, por exemplo, em (Hall et al., 2011; Kohavi, 1995; Salzberg & Fayyad, 1997), o passo seguinte foi a separação do corpus em dois conjuntos disjuntos. Um de teste, com 10% das frases, e outro de treino, com os restantes 90%. Obedecendo a esta métrica, foram gerados 10 grupos distintos. Como as diversas frases da linguagem de saída do corpus foram obtidas, inicialmente, por replicação, se se limitasse a fazer uma separação por simples sorteio aleatório, corria-se o risco de, num mesmo grupo, existir uma maior preponderância de frases com a mesma 'semente'. Por este motivo, as frases foram separadas nos 10 grupos da seguinte forma:

1. Mantendo a sequência inicial das frases, a cada frase foi atribuída uma referência, correspondente a uma letra do alfabeto. Foram utilizadas as letras A a J, identificando cada letra um grupo.
2. A sequência A a J foi atribuída sequencialmente, renovando-se continuamente. Desta forma foi garantido que as frases obtidas a partir de uma dada 'semente', ficam distribuídas por grupos totalmente distintos.

Linguagem interna	**Frase correspondente**
pessoa32n saudacao_0 pessoa0a medicamento21 tipo0 tomar0 cor00 dose0 freqtoma00	Helena pode tomar agora o Seretaide.
pessoa0n saudacao_0 pessoa0a medicamento14 tipo1 tomar2 cor00 dose4 freqtoma02	Vai-se deitar então tome quatro comprimidos Primperan.
pessoa40n saudacao_0 pessoa0a medicamento0 tipo1 tomar2 cor03 dose0 freqtoma04	Ao almoço toma o comprimido branco Leonardo.
pessoa0n saudacao_m pessoa12a medicamento0 tipo8 tomar3 cor00 dose0 freqtoma02	Antes de deitar senhor Lima não se esqueça da bomba de inalação.
pessoa17n saudacao_f pessoa0a medicamento0 tipo4 tomar2 cor04 dose0 freqtoma02	Antes de deitar faça a toma das gotas amarelas dona Cristina.
pessoa0n saudacao_0 pessoa0a medicamento3 tipo1 tomar2 cor10 dose4 freqtoma05	É hora de jantar tome os quatro comprimidos laranja do Ibuprofeno.
pessoa36n saudacao_0 pessoa0a medicamento19 tipo8 tomar3 cor00 dose0 freqtoma01	São horas de acordar e de colocar a bomba de inalação Nasomet daqui a três horas João terá de colocar de novo.
pessoa21n saudacao_f pessoa0a medicamento2 tipo4 tomar2 cor00 dose5 freqtoma01	Dona Elisabete assim que acordar deve tomar cinco gotas de Clorocil.
pessoa37n saudacao_0 pessoa0a medicamento23 tipo4 tomar2 cor00 dose4 freqtoma05	Está na hora de jantar Jorge não esqueça de tomar as quatro gotas de Guttalax.
pessoa78n saudacao_f pessoa0a medicamento3 tipo1 tomar2 cor00 dose3 freqtoma04	Dona Teresinha está na hora de almoço tome os três comprimidos Ibuprofeno.

Tabela 3: Parte do corpus alinhado utilizado nas experiências (as 10 primeiras linhas do corpus de treino A).

Adriana podes tomar agora o Clorocil
D. Teresa antes de deitar coloque as gotas Clorocil
NOME podes tomar agora o MEDICAMENTO
FORMA_TRATAMENTO NOME antes de deitar coloque as TIPO MEDICAMENTO

Tabela 4: Duas frases recolhidas no corpus original e correspondente adaptação com os *tokens*.

Token	**Correspondência**
NOME	nome do destinatário do medicamento
APELIDO	apelido do destinatário do medicamento
FORMA_TRATAMENTO	saudação ao destinatário do medicamento (corresponde a Sr., Sra., D., etc.)
MEDICAMENTO	nome do medicamento
QUANTIDADE	quantidade a tomar do medicamento
TIPO	tipo de medicamento (comprimidos, gotas, etc.)
TEMPO	hora do dia em que o medicamento deveria ser tomado
COR	cor do medicamento

Tabela 5: *Tokens* e respectivas correspondências.

3. Após a classificação das frases, o corpus foi ordenado pela sua referência (letra A a J), constituindo-se assim 10 grupos disjuntos.

4. Ao fazer a separação do corpus desta forma foi garantida a aleatoriedade da constituição de cada grupo. A produção inicial de cada uma das frases 'semente' é independente. A replicação das 126 frases para as 643 finais é aleatória, já que cada frase foi replicada, aleatoriamente, entre 3 e 7 vezes. A substituição dos *tokens* por valores concretos foi efectuada com valores escolhidos aleatoriamente. Na distribuição de frases por grupos, todas as frases de cada grupo têm uma 'semente' distinta.

Concretizada a separação do corpus em 10 grupos disjuntos, foram constituídos 10 conjuntos de teste e de treino. Cada conjunto, igualmente denominado por uma letra de A a J e sufixado respectivamente por "-teste" e "-treino", corresponde ao seguinte: o conjunto de teste tem o nome igual ao do grupo obtido, como acima explicado; O conjunto de treino contém as frases de todos os restantes grupos.

4.4 Algumas estatísticas

Alguma informação estatística acerca da 'linguagem de saída' do corpus pode ser encontrada na Tabela 6.

Número de frases	643
Número de palavras	7212
Número médio de palavras/frase	11
Número máximo de palavras/frase	30
Número mínimo de palavras/frase	4

Tabela 6: Alguma informação estatística relativa à 'linguagem de saída' do corpus.

5 Sistemas Desenvolvidos

Foram desenvolvidas duas variantes do sistema, aproveitando as duas variantes principais dos sistemas de tradução automática: baseadas em sintagmas (*phrase-based*) e usando informação sintáctica (*tree-based*).

5.1 Sistemas baseados em sintagmas (*phrase-based*)

Para a execução desta primeira experiência, após a recolha do corpus, respectiva expansão e segmentação procedeu-se à operação de treino. Foram executados os diversos procedimentos, conforme prescrito em (MOSES, 2014a). O Moses dispõe de diversos *scripts* especialmente preparados para a execução de cada uma das fases. Para cada um dos 10 conjuntos de treino, foi efectuada uma operação de treino, e teste, totalmente independente. Cada conjunto foi submetido aos mesmos procedimentos, executados pela mesma ordem.

A primeira tarefa consistiu na preparação do corpus. Em primeiro lugar foi efectuada a *tokenização* do corpus. Esta operação consiste em separar, com um espaço em branco, antes e depois, cada um dos elementos que constituem cada uma das frases do corpus. Aqui, por "elemento" entende-se cada palavra e sinal de pontuação existentes na frase. Na 'linguagem de saída' este procedimento foi realizado com recurso ao script 'tokenizer.perl'. Na 'linguagem de entrada' não foi necessário efectuar esta tarefa, pois da forma como ela foi criada, todos os elementos de cada frase estão naturalmente *tokenizados*.

A fase de limpeza foi executada, contudo não teve efeitos práticos. Efectivamente, esta fase destina-se a eliminar as frases mal formadas e com tamanho excessivo. Considera-se tamanho excessivo uma frase com mais de 80 palavras. No nosso corpus, todas as frases têm dimensão inferior a esse limite e encontram-se bem formadas e devidamente alinhadas.

A segunda tarefa consistiu no treino do modelo de linguagem. Nesta fase, são criadas as ferramentas intermédias que vão assistir o Moses na realização do treino do sistema de tradução. Este modelo intermédio destina-se a assegurar uma geração fluente do texto produzido, sendo por isso efectuada sobre a 'linguagem de saída'. Neste passo, foi utilizado o IRSTLM (IRSTLM, 2011).

Neste mesmo passo, é, também, definido o parâmetro *ngram*. Por defeito, tem o valor 3, o que significa que o modelo irá efectuar agrupamentos de palavras até 3 elementos. Estes agrupamentos serão posteriormente utilizados na criação dos textos de saída. São essencialmente executados 3 passos. Primeiro, cada frase é prefixada com o termo <s>e sufixada com o correspondente termo </s>. Depois, o modelo da linguagem é construído. Por último, este modelo é compilado.

A terceira tarefa consistiu no treino do sistema de tradução. Concluídas as duas primeiras tarefas, estão reunidas as condições para se efectuar o treino do sistema de tradução. Esta tarefa recorre ao modelo da linguagem, gerado na tarefa anterior, e ao software GIZA++ (Och, 2011). Aqui são gerados, entre outros, os ficheiros 'moses.ini' e 'phase-table.gz' necessários à configuração e utilização do Moses.

A última tarefa consiste no teste, e uso, do modelo treinado para se gerarem os textos pretendidos.

5.2 Sistemas usando informação sintáctica (*tree-based)*

Na execução com o modo *tree-based*, foram utilizados os mesmos conjuntos, já referidos, tendo sido realizadas experiências independentes. A grande diferença entre o treino anterior e este treino centra-se na construção da árvore que representa a 'linguagem de saída'. Posteriormente,

é o ficheiro com a árvore que é utilizado no treino do sistema de tradução.

A produção da árvore, que permitiu classificar cada palavra, de cada frase, em termos da sua função morfossintáctica (nome, verbo, adjectivo, etc.), revelou-se uma tarefa difícil de superar.

A principal dificuldade resultou da escolha do *parser* a utilizar. O nossos principais requisitos eram: (1) produzir uma classificação que pudesse ser facilmente adaptada para utilização no MOSES e que fosse compatível com as ferramentas de manipulação de "árvores" deste sistema; (2) possibilidade de integração do *parser* no nosso sistema; (3) utilização gratuita.

Consideradas estas exigências, foram instalados e testados diversos *parsers*, nomeadamente: o Palavras[1], o Freeling[2], o Tycho Brahe[3], o TreeTagger[4], o Turbo Semantic Parser[5] e o Stanford Parser[6].

A escolha recaiu sobre este último, com a adaptação efectuada pelo grupo LX-CENTER (Language Resources and Technology for Portuguese), da Universidade de Lisboa – Portugal (Branco & Silva, 2004). Apesar das limitações verificadas, foi, no entanto, o *parser* que melhor correspondeu aos nossos requisitos.

6 Resultados

Nesta secção, apresentam-se exemplos representativos das capacidades de geração dos sistemas desenvolvidos. Segue-se a apresentação de uma avaliação formal, usando métricas comuns na área e avaliação por humanos. A eventual influência da forma de divisão do corpus (em treino e teste) é também avaliada. No final, apresenta-se alguma informação sobre a integração em curso na aplicação para *Smartphones* denominada Assistente de Medicação.

6.1 Exemplos

Na Tabela 7, apresentam-se vários exemplos seleccionados de forma a ilustrar os vários tipos de resultados obtidos. Pretende-se familiarizar o leitor com o que de facto foi possível obter, usando os dois tipos de sistemas.

No topo da tabela, com os números 1 a 3, apresentam-se alguns exemplos de representação interna e as correspondentes frases gerada por um dos sistemas baseado em sintagmas (concretamente o treinado com a parte A do corpus). As frases geradas têm qualidade bastante diversa, sendo uma intelegível, outra considerada aceitável e a restante considerada correta.

Na segunda parte da tabela (números 4 a 7) apresenta-se o alinhamento entre as frases criadas pelos humanos e as criadas pelo sistema (independentemente do seu tipo) para uma mesma entrada. As frases são, aqui, apresentadas em minúsculas para que seja possível evidenciar as suas diferenças, como a seguir explicado. Para uma certa frase de entrada, quando uma frase é gerada pelo nosso sistema, ela pode apresentar (e normalmente apresenta) diferenças com a frase de treino. Essas diferenças podem corresponder a adições de novas palavras, supressão de palavras ou troca de posição de palavras dentro da frase. Aqui, utilizamos a marcação "***" para evidenciar a adição ou supressão de palavras e as maiúsculas para assinalar a troca de posição de palavras. Estas ocorrências surgem apenas na frase gerada. Quando as representamos na frase escrita por humanos, pretendemos, apenas, tornar mais evidente as diferenças entre as duas frases.

Os exemplos 6 e 7 apresentam frases geradas, consideradas inteligíveis e boas em termos de naturalidade, mas que são completamente diferentes das produzidas por humanos. Este tipo de frases constitui um grande desafio para a avaliação, sendo normalmente consideradas como erros pelas métricas automáticas de avaliação. Tendo em conta estas ocorrências, tivemos necessidade de reconsiderar a nossa opção inicial de apenas aplicar avaliação automática, avaliando uma parte dos resultados por humanos.

Na terceira parte da tabela, exemplifica-se a diferença entre os resultados obtidos pelos 2 tipos de sistema. Para cada um dos subconjuntos, foi utilizado o mesmo vector de entrada, para garantir a comparabilidade das frases.

6.2 Avaliação comparativa dos dois tipos de sistemas

6.2.1 Método

Tendo em conta os objectivos principais de ter informação sobre o desempenho absoluto e relativo dos dois tipos de sistemas criados, começou-se por treinar 10 sistemas para cada um dos tipos, adoptando-se os valores por defeito para a generalidade dos processos (ou seja o valor de 3 para o

[1] http://beta.visl.sdu.dk/contact.html
[2] http://nlp.lsi.upc.edu/freeling/
[3] http://www.tycho.iel.unicamp.br/~tycho/apps/dbparser-files/
[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
[5] http://labs.priberam.com/Resources/TurboSemanticParser.aspx
[6] http://nlp.stanford.edu/software/lex-parser.shtml

Exemplos mostrando a linguagem de entrada e a frase gerada	
Num	**Exemplo**
1	pessoa45n saudacao0 pessoa0a medicamento17 tipo0 tomar0 cor00 dose0 freqtoma00 Luís Pulmicort de tomar agora o
2	pessoa61n saudacaom pessoa0a medicamento4 tipo0 tomar0 cor00 dose2 freqtoma00 senhor Paulo tome dois comprimidos de Salazopirina
3	pessoa49n saudacao0 pessoa0a medicamento0 tipo2 tomar2 cor09 dose0 freqtoma10 Marcelo aplique ao tomar a cápsula branca e azul de dez horas em dez horas
Exemplos de saída dos sistemas (S) alinhados com frases produzidas por humanos (H)	
4	H: dona denise assim que se levantar não se esqueça de tomar OS COMPRIMIDOS nicotibine S: dona denise assim que se levantar não se esqueça de tomar O COMPRIMIDO nicotibine
5	H: DEVE TOMAR AGORA ao acordar *** a bomba de inalação DE pulmicort AUGUSTO S: *** *** AUGUSTO ao acordar APLIQUE a bomba de inalação *** pulmicort ***
Exemplos de geração muito diferentes da frase de teste, mas aceitáveis	
6	H: É HORA DE ALMOÇAR marcos não se esqueça de tomar *** quatro gotas de guttalax *** *** S: *** *** *** *** marcos não se esqueça de tomar AS quatro gotas de guttalax AO ALMOÇO
7	H: *** *** É MEIO-DIA TOME AS três gotas de zaditen *** PATRÍCIA S: PATRÍCIA NÃO SE ESQUEÇA DE TOMAR três gotas de zaditen AO MEIO-DIA
Saídas produzidas pelos 2 sistemas, para uma mesma entrada (F identifica o baseado em sintagmas, S2T identifica o baseado em informação sintáctica)	
8	F: Patrícia não se esqueça de tomar três gotas de Zaditen ao meio-dia S2T: Patrícia ao Zaditen gotas tome de três meio-dia
9	F: Senhora Carvalho após o seu almoço tome cinco comprimidos de Duphaston S2T: Senhora Carvalho Duphaston comprimido branco cinco almoço

Tabela 7: Exemplos de saídas dos dois tipos de sistemas.

ngram). Depois de treinados, os sistemas foram avaliados, primeiro com o corpus de treino e, após verificado o bom funcionamento do sistema, com o conjunto de teste correspondente. Adoptaram-se para a avaliação, os parâmetros BLEU (Papineni et al., 2002) e Meteor (Denkowski & Lavie, 2014, 2011), seguindo, por exemplo (Langner, 2010).

Complementarmente, foi realizada uma avaliação por humanos. As frases avaliadas foram escolhidas de entre as geradas pelos 2 sistemas. Todas estas frases foram obtidas com o mesmo conjunto de teste — o conjunto F. Depois de escolhidas, foram ordenadas aleatoriamente e avaliadas em termos de inteligibilidade, estrutura da frase e qualidade global. Para simplificar a tarefa dos avaliadores, as respostas possíveis para inteligibilidade e estrutura foram reduzidas a apenas 3 opções. Informação concreta sobre as questões e as opções de resposta encontram-se na Tabela 8. Participaram na avaliação 11 pessoas, com características muito diferentes, quer em termos de idade (variaram entre os 16 anos e os 58 anos), quer em termos de formação e actividade profissional (e.g. estudantes, assistentes administrativos e professores).

	Questão	**Opções de resposta**
Inteligib.	Percebe-se ?	0 = Não 1 = Mais ou menos 2 = Sim
Estrutura	Estrutura da Frase	0 = Má (vários problemas) 1 = Mais ou menos 2 = Boa
Qualidade	Qualidade Geral ?	de 1 a 5, onde 1 = Má 5 = Excelente

Tabela 8: Informação sobre as questões e opções de resposta utilizadas na avaliação das frases geradas por humanos.

No caso dos sistemas baseados em sintaxe, foi feita uma experiência relativa ao efeito do peso atribuído ao modelo de linguagem no processo de *decoding*. Foram experimentados vários pesos, usando o corpus de treino, tendo-se chegado à conclusão de que existia um efeito muito positivo nas métricas BLEU e Meteor quando esse peso era bastante superior ao valor por omissão. Tendo em conta este resultado, este novo valor (de 10) foi adoptado para as avaliações de todos os sistemas baseados em sintaxe.

6.2.2 Resultados da avaliação automática

Os resultados obtidos em termos de BLEU e Meteor para os dois tipos de sistemas, são apresentados nas Figuras 4 e 5, sob a forma de média e intervalo de confiança a 95%.

Em termos de BLEU, na Figura 4, o sistema baseado em sintagmas obtém um melhor desempenho médio, em todos os parâmetros, excepto para o parâmetro Bleu 1 – associado a unigramas –, sendo mesmo o desempenho significativamente superior (visível pela não sobreposição dos intervalos de confiança) para o parâmetro global (BLEU na figura) e para o Bleu 2. O melhor valor de BLEU obtido foi de 0,245.

Os resultados anteriores são, de um modo geral, confirmados pelo Meteor (ver Figura 5). Neste caso, todos os parâmetros são piores para o sistema baseado em sintaxe, sendo de destacar as diferenças em termos de 'Recall', 'Penalização Devida a Fragmentação' e na 'Score Final'. Enquanto ambos os sistemas são capazes de um desempenho similar em termos de precisão, acertando em geral nas palavras que seleccionam para a frase, o sistema baseado em sintaxe falha muito mais na inclusão de palavras que deviam constituir a frase.

Os resultados obtidos pelo sistema baseado em sintaxe estão certamente relacionados com a qualidade da classificação sintáctica efectuada pelo *parser* escolhido. A Figura 6 apresenta três exemplos de frases com problemas na classificação. São evidentes erros na classificação de verbos, nomes e artigos. O problema mais recorrente é a deficiente classificação dos nomes dos medicamentos.

```
(ROOT (S
   (NP (N Daniel))
   (VP (V deve) (VP (V' (V tomar) (ADV agora))
   (NP (NP (ART o) (N comprimido))
   (S (VP (V Duphaston) (NP (ART uma) (N' (N vez) (CP (NP (REL que))
   (S (VP (V são) (NP (CARD vinte) (N horas)))))))))))))

(ROOT (S
   (NP (NP (N' (N No) (N fim) (N do) (N jantar)))
   (NP (N' (N dona) (N Inês))))
   (VP (V aplique) (NP (ART a) (N' (N pomada) (A Fucithalmic))))))

(ROOT (S
   (S (CONJ Quando) (S (NP (CL se)) (VP (V levantar)
      (NP (N' (N senhora)
      (N Pereira) (N tome))))))
   (S (NP (ART as) (N' (CARD duas) (N gotas)))
      (VP (V' (V Neo-Sinedrina)
      (ADV depois))
      (NP (N' (N repita) (N ao) (N almoço)))))))
```

Figura 6: Exemplos de erros na classificação de palavras, originado pelo *parser* Stanford.

A pequena extensão dos intervalos de confiança mostra uma pequena variação dos resultados com a divisão do corpus utilizada para teste.

Da conjugação dos resultados obtidos nas duas métricas de avaliação, resulta claro um melhor desempenho do sistema baseado em sintagmas para esta tarefa.

6.2.3 Resultados da avaliação por humanos

Os resultados da avaliação por humanos são resumidos nas Figuras 7 a 9. Em cada figura, são apresentadas as contagens de cada uma das opções de resposta, em confronto com os resultados dos dois tipos de sistema.

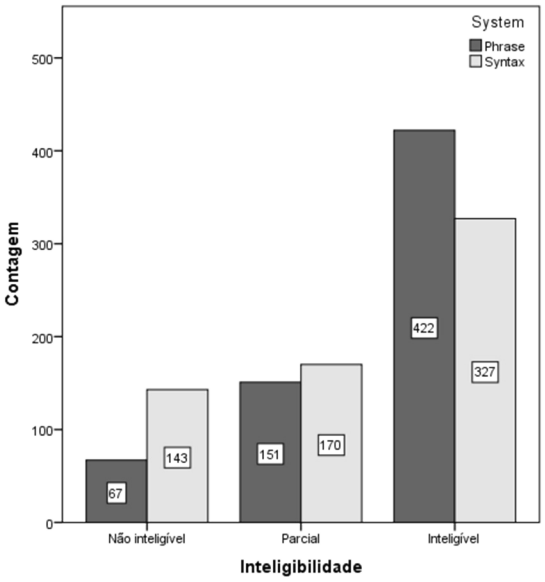


Figura 7: Distribuição das respostas da questão relativa à inteligibilidade das frases. As barras mais escuras referem-se ao sistema baseado em sintagmas.

Em termos de inteligibilidade, o sistema baseado em sintagmas obteve um maior número de respostas, indicando frases inteligíveis (uma diferença de 95 respostas, o que significa uma média de mais 8,7 respostas positivas em termos de inteligibilidade por avaliador). Sendo a diferença entre os dois sistemas para avaliações indicando inteligibilidade parcial baixa, este pior desempenho do sistema baseado em sintaxe reflecte-se no maior número de frases avaliadas como não inteligíveis. Enquanto o sistema baseado em sintagmas apresenta uma média de 6 frases não inteligíveis por avaliador, o sistema baseado em sintaxe apresenta um valor médio superior ao dobro (13). Tendo em conta que foram avaliadas 64 frases para cada sistema, estes valores correspondem a, respectivamente, 9 e 20 % de frases não-inteligíveis. Do lado positivo, foram consideras como inteligíveis, por cada avaliador, uma média de 60 % e 46 %.

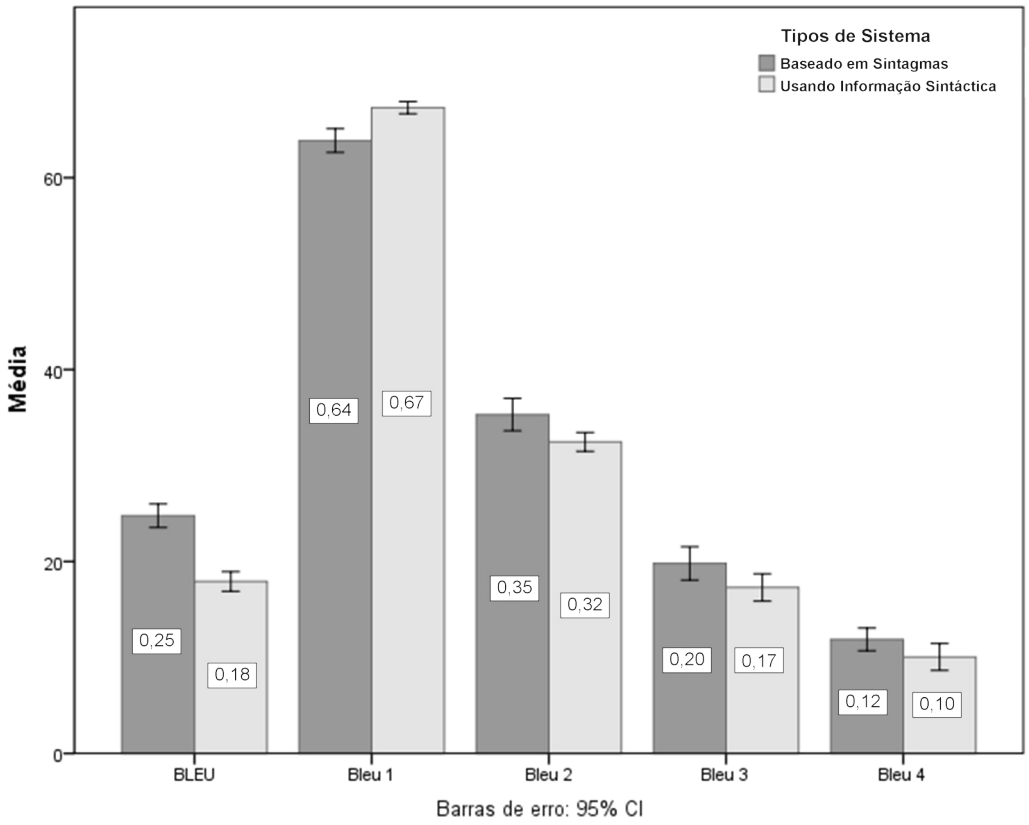


Figura 4: Resultados da avaliação baseada no BLEU.

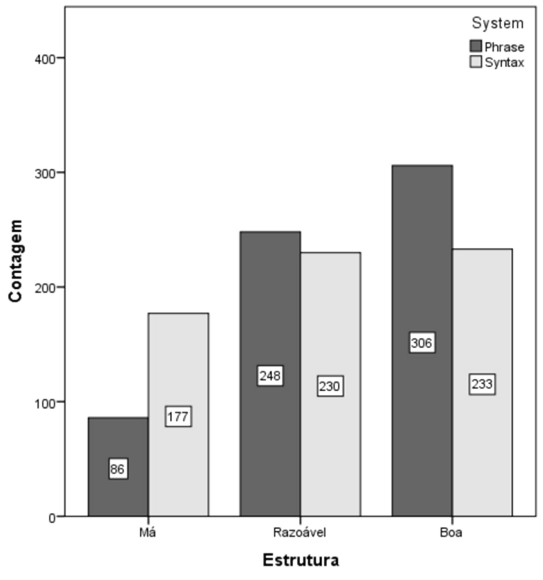


Figura 8: Resultados da avaliação da estrutura das frases. As barras mais escuras correspondem ao sistema baseado em sintagmas.

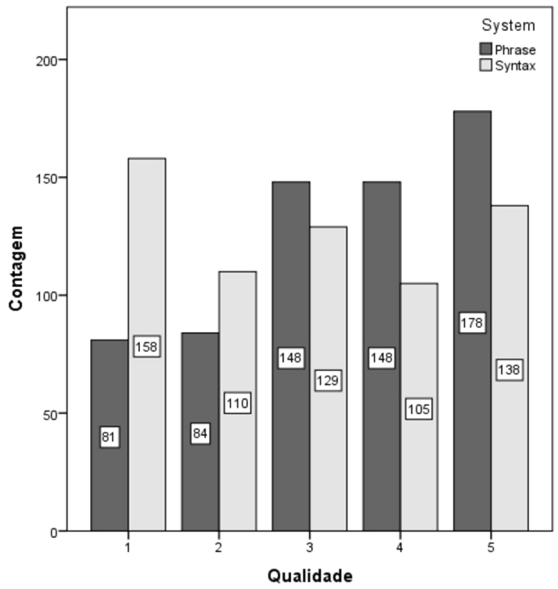


Figura 9: Distribuição das respostas da questão relativa à qualidade geral das frases. As barras mais escuras referem-se ao sistema baseado em sintagmas.

Em termos de estrutura das frases (Figura 8), mantém-se o melhor desempenho do sistema baseado em sintagmas, com um valor próximo do dobro de frases avaliadas como tendo vários problemas de estrutura. Mais uma vez, as grandes diferenças ocorrem nos extremos (má estrutura e boa estrutura).

Na avaliação geral da qualidade (Figura 9), o sistema baseado em sintagmas apresenta número de avaliações superiores ao baseado em sintaxe para os valores geralmente interpretados como indicando uma avaliação positiva (iguais ou su-

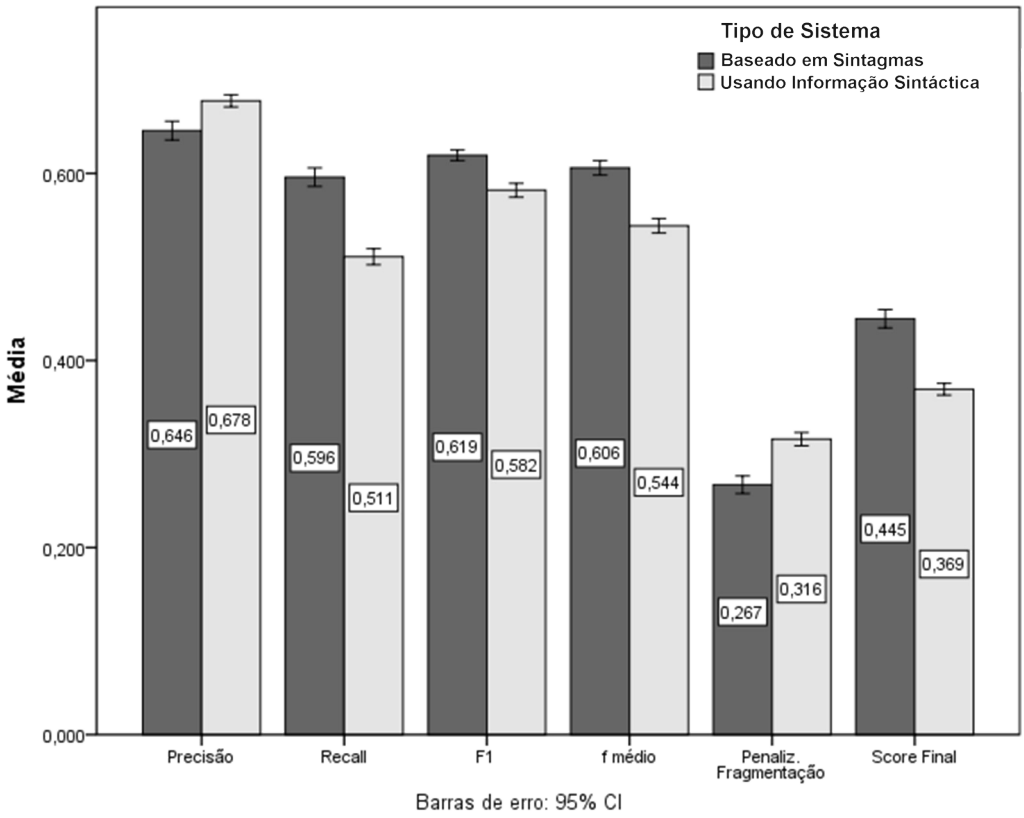


Figura 5: Resultados da avaliação usando o Meteor.

periores a 3). O sistema baseado em sintaxe obteve um valor muito superior de avaliações com o valor mais baixo da escala (1). Ambos os sistemas apresentam uma distribuição das classificações pelos 5 valores da escala, com tendência, no caso do baseado em sintagmas, para uma preferência pelos valores entre 3 e 5. É importante destacar que, para o melhor sistema, foram, em média, avaliadas como boas ou excelentes 46 % e como excelentes 25 % das frases.

6.3 Efeito da forma de divisão do corpus

De forma a descartar um possível efeito do modo como foi dividido inicialmente o corpus (ver Secção 4.3), foi criada uma nova experiência, onde o corpus foi dividido aleatoriamente em 10 novos conjuntos, com número similar de frases. Nesta experiência, tendo em consideração os melhores resultados inicialmente obtidos, foi decidido utilizar a versão baseada em sintagmas. Estes novos conjuntos foram treinados e testados de forma equivalente à experiência anterior. A Figura 10 evidencia as diferenças obtidas nas duas experiências.

É notório para os dois conjuntos de métricas que o desempenho, avaliado nos 10 conjuntos de teste, é superior quando utilizado o método de divisão proposto neste artigo. Esta diferença é estatisticamente significativa para um nível de confiança de 5 % (os intervalos de confiança, CI, a 95 % não se sobrepõem).

Tendo em conta os resultados superiores com a divisão não-aleatória do corpus, decidiu-se pela não realização de mais treinos e testes com as divisões obtidas pelo processo de divisão aleatória.

6.4 Variabilidade das respostas do sistema

Um bom sistema baseado em *templates* pode produzir interacção com o ser humano com muita naturalidade, variabilidade e qualidade. Contudo, para que tal aconteça, é necessário efectuar um grande investimento, seja em tempo, seja em recursos.

Experiências efectuadas com o nosso sistema mostraram que, para um vector de entrada similar, são produzidas respostas corretas e distintas, que não fazem parte do corpus inicial. Este tipo de geração permite criar, com facilidade, novas respostas, proporcionando variabilidade na interacção com o seu utilizador. A Tabela 9 apresenta dois exemplos onde, para entradas similares, obtemos respostas distintas. Nos exemplos 1 e 2, apenas varia o nome da pessoa referenciada. Nos exemplos 3, 4 e 5, apenas varia o nome do medicamento, e naturalmente, a sua forma e tipo de toma.

Num	Exemplo
1	pessoa18n saudacao_m pessoa0a medicamento24 tipo3 tomar1 cor00 dose0 freqtoma20 Senhor Daniel aplique a pomada Fucithalmic que são vinte horas
2	pessoa18n saudacao_m pessoa5a medicamento24 tipo3 tomar1 cor00 dose0 freqtoma20 Senhor Daniel Costa deve aplicar a pomada Fucithalmic que são vinte horas
3	pessoa18n saudacao_m pessoa0a medicamento24 tipo3 tomar1 cor00 dose0 freqtoma20 Senhor Daniel aplique a pomada Fucithalmic que são vinte horas
4	pessoa18n saudacao_m pessoa0a medicamento3 tipo1 tomar2 cor00 dose0 freqtoma20 Senhor Daniel não se esqueça de tomar o comprimido Ibuprofeno são vinte horas
5	pessoa18n saudacao_m pessoa0a medicamento12 tipo1 tomar2 cor00 dose0 freqtoma20 Senhor Daniel tome o comprimido de Nicotibine são vinte horas

Tabela 9: Exemplo de geração de respostas, associadas a entradas similares.

6.5 Primeiras integrações

Uma versão inicial dos sistemas aqui apresentados, baseada em sintagmas, foi alvo de uma primeira integração numa aplicação real para assistência à toma de medicamentos por idosos. Este sistema, para Smartphones, e desenvolvido no âmbito do projecto *Smartphones for Seniors*, foi descrito e avaliado em (Teixeira et al., 2013a; Ferreira et al., 2013a,b)

7 Discussão

Os resultados obtidos pelo melhor sistema, em termos de inteligibilidade e qualidade das frases geradas, indicam que a abordagem adoptada e os sistemas desenvolvidos conseguem gerar frases de qualidade similar às produzidas por humanos. Contudo quando falham, as frases geradas podem ser completamente ininteligíveis.

Estes resultados estão de acordo com os relatados por Langner (Langner, 2010) para o sistema Mountain. No Mountain, foi apenas utilizada a variante de desenvolvimento por sintagmas (*phrase-based*), pelo que o nosso comentário se refere apenas a este tipo de geração. A Tabela 10 apresenta os resultados por ele obtidos. As variantes ao sistema base (*Rating >1 . . . Rating >4*) correspondem a experiências efectuados por Langner. Essas experiências foram motivadas pela qualidade do corpus utilizado no Mountain. Cada experiência corresponde, assim, a retiradas sucessivas, ao corpus, das frases classificadas, por avaliadores humanos, como inadequadas. Verifica-se que, até certo ponto, quanto menos frases consideradas 'más' estiverem presentes no corpus, melhor o índice BLEU.

Da comparação entre os dados da Figura 4 e da Tabela 10 resulta que o nosso sistema apresenta valores ligeiramente superiores ao Mountain (qualquer que seja a sua versão), quando se comparam os dados referentes aos Bleu1, Bleu2, Bleu3 e Bleu4.

A Tabela 11 apresenta os dados da avaliação do sistema Mountain, nas mesmas condições referidas atrás, segundo o método Meteor. Comparando os dados desta figura com os dados da Figura 5, verifica-se que, à semelhança do observado para o método BLEU, os resultados do nosso sistema são ligeiramente melhores, nos seus diversos indicadores.

Estas observações permitem-nos concluir que, apesar das limitações do nosso sistema, o seu desempenho é satisfatório e está alinhado com o desempenho de sistemas similares.

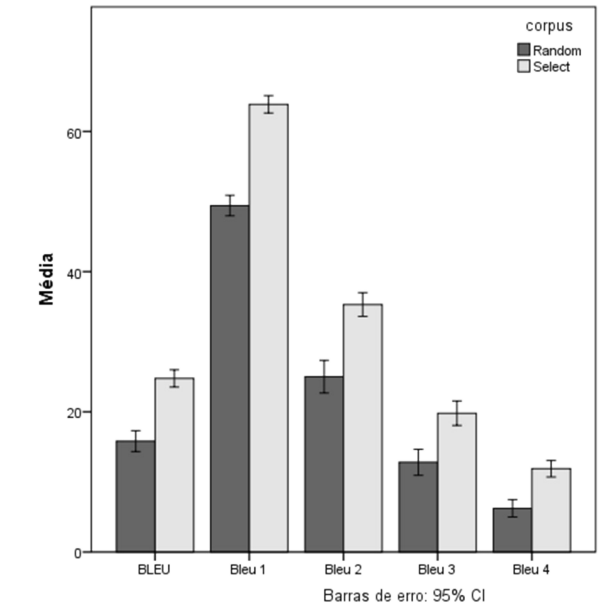
System	1-gram	2-gram	3-gram	4-gram	5-gram
Baseline	0.3198	0.1022	0.0525	0.0300	0.0202
Rating > 1	0.4376	0.1729	0.1079	0.0746	0.0597
Rating > 2	**0.4491**	**0.1919**	0.1169	0.0872	0.0747
Rating > 3	**0.4742**	**0.1963**	0.1212	0.0866	0.0722
Rating > 4	**0.4596**	0.1762	0.1023	0.0693	0.0611

Tabela 10: Avaliação do sistema Mountain (Langner, 2010, pag. 73), pelo método BLEU.

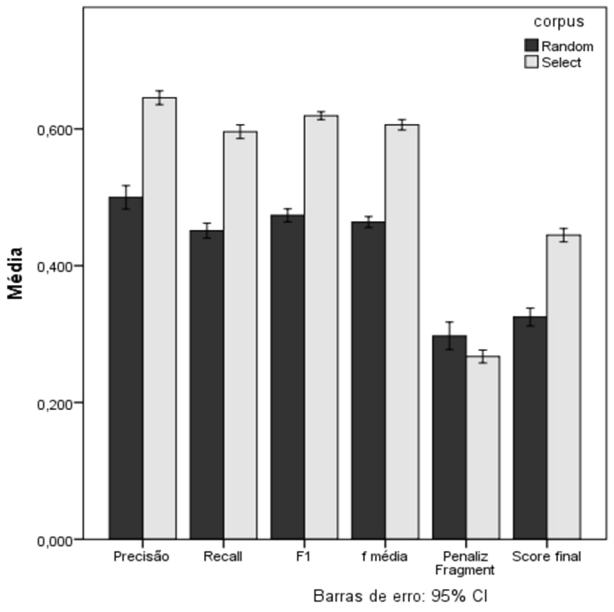
System	Precision	Recall	F_1	Total
Baseline	0.4225	0.2013	0.2727	0.1950
Rating > 1	0.4489	0.2097	0.2859	0.2028
Rating > 2	0.4533	**0.2248**	**0.3009**	**0.2218**
Rating > 3	**0.4834**	0.2148	0.2974	0.2146
Rating > 4	0.4481	0.2030	0.2794	0.1971

Tabela 11: Avaliação do sistema Mountain (Langner, 2010, pag. 75), pelo método Meteor.

Um dos objectivos que pretendemos alcançar, é a geração de um sistema em que o esforço necessário para a obtenção de corpora seja pequeno. Tendo em conta este requisito, os sistemas foram capazes de gerar um conjunto interessante de frases, apesar de terem sido treinados apenas com um pequeno corpus, criado tendo por base pouco mais de 100 frases efectivamente escritas por humanos. Este facto demonstra não só o potencial para melhoramento, através do aumento do corpus, mas também o potencial para se criar sistemas minimamente úteis com muito pouco investi-



a) BLEU



b) METEOR

Figura 10: Resultados obtidos com os dois métodos de divisão do corpus utilizados. Os resultados referem-se apenas aos sistemas baseados em sintagmas.

mento na criação de corpora. O melhor desempenho do sistema mais simples (baseado em sintagmas) não deve ser interpretado como um indício da desadequação do sistema usando informação sintáctica. Como potenciais causas deste desempenho inferior temos: (1) reduzido tamanho do corpus, que pode ser insuficiente para treinar adequadamente os modelos, certamente mais exigentes; (2) efeito negativo dos erros de análise sintáctica. Quanto ao segundo problema, estamos convictos que, com a utilização de um *parser* que classifique melhor as diversas palavras das frases, quanto à sua função sintáctica, o desempenho do sistema baseado em sintaxe melhorará.

Por outro lado, o facto de os dois sistemas gerarem muitas vezes frases muito diferentes pode ser explorado na criação de um sistema em que os resultados de ambos sejam objecto de um processo de avaliação e selecção da melhor frase. Acreditamos que o desempenho de ambos os sistemas é passível de ser melhorado, através de um processo de afinação (*tuning*) dos muitos parâmetros dos modelos, usando um corpus de validação.

Um outro aspecto positivo a destacar é o aumento do desempenho obtido, através de um método não-aleatório de divisão do corpus em treino e teste. uma vez que se aplicou um processo de expansão de um conjunto base de frases na criação do corpus, o método usual de divisão aleatória não é o mais adequado, não garantindo que exemplos derivados de uma mesma frase fiquem devidamente divididos entre os conjuntos de treino e teste. O método proposto evita que se reduza em demasia a presença, no conjunto de treino, de exemplos resultantes de uma mesma frase base.

A grande limitação dos sistemas criados é a sua imprevisibilidade em termos de qualidade dos resultados. A avaliação, como esperado, revelou-se uma tarefa bastante complicada, com as métricas automáticas a apresentar grandes dificuldades em fornecer informação adequada. Apenas com a utilização combinada de 2 métricas e avaliação por humanos foi possível ter uma visão minimamente clara sobre o desempenho dos sistemas. Recentemente, foi apresentada (Pereira et al., 2015) uma primeira extensão a este sistema, onde é feita uma proposta de avaliação automática da qualidade, recorrendo à extração e análise de características (*features*) sobre as frases geradas.

8 Conclusão

Motivados pela crescente necessidade de transmitir, em português, informação gerada por sistemas computacionais cada vez mais sofisticados e omnipresentes, neste artigo apresenta-se uma primeira experiência para a língua portuguesa na área da geração de frases a partir de dados referentes a planos de medicação. O sistema adoptado baseia-se na utilização de tradução automática e inspira-se em trabalhos recentes, como o sistema Mountain. Foram desenvolvidos e avaliados comparativamente dois tipos de sistemas

de tradução, um baseado em sintagmas e outro usando informação sobre a sintaxe. Os resultados da avaliação, englobando avaliação automática e avaliação por humanos, mostram que o sistema baseado em sintagmas obteve o melhor desempenho. Este tipo de sistemas é capaz de gerar uma boa percentagem de frases inteligíveis ou minimamente inteligíveis (menos de 10 % de frases não inteligíveis), com percentagem interessante das frases geradas a obter avaliações de qualidade geral de nível bom ou superior.

Para divisão do corpus para o treino de teste de 10 sistemas (*10-fold Cross-Validation*) foi desenvolvido um processo alternativo à usual divisão aleatória, que se revelou capaz de contribuir para um melhor desempenho dos sistemas testados.

8.1 Trabalho Futuro

Uma continuação óbvia do trabalho aqui apresentado passa pelo aumento do corpus. Para esta primeira experiência considerou-se importante não investir muitos recursos na criação de um corpus extenso, mas é importante investigar o efeito do tamanho do corpus no desempenho deste tipo de sistema.

Não produzindo os sistemas desenvolvidos uma percentagem de frases inteligíveis próxima dos 100 %, de forma a tornar estes sistemas utilizáveis numa aplicação real, como é nosso objectivo, torna-se necessário desenvolver um módulo que disponibilize uma estimativa da inteligibilidade e naturalidade das frases geradas. Com essa informação, será possível criar um sistema híbrido que recorra a *templates* quando essa estimativa aponte para uma frase de baixa qualidade e em que a inteligibilidade esteja comprometida.

Nesta fase do nosso trabalho, a nossa principal preocupação, foi determinar se as frases geradas eram facilmente compreendidas por qualquer pessoa. A etapa seguinte será avaliar a sua adequabilidade, com recurso a profissionais da área do tema da aplicação.

Por último, mas não menos importante, interessa-nos explorar formas rápidas de aplicar este tipo de abordagens a outras aplicações. Interessa-nos, também, reforçar a multimodalidade da aplicação. Nomeadamente, com recurso à síntese de fala e/ou com recurso a imagens. Só desta forma será possível complementar a informação prestada e eliminar algumas ambiguidades que eventualmente existam.

Agradecimentos

Os autores agradecem a todos os que contribuíram para a criação do corpus e a todos os que participaram na avaliação das frases. Um agradecimento especial ao Mário Rodrigues pela ajuda na obtenção e utilização dos *parsers* sintácticos para o português.

Um último agradecimento para os vários revisores deste artigo, que em muito contribuíram para a sua evolução.

Referências

Agência Lusa. 2010. Falantes de português irão aumentar para 335 milhões em 2050. *Público (Online, verificado em 23/07/2015)* http://www.publico.pt/culturaipsilon/noticia/falantes-de-portugues-irao-aumentar-para-335-milhoes-em-2050-1429372.

Alves, Lúcia Vinheiras. 2011. Português é terceira língua mais falada no mundo. Online, verificado em 23/07/2015. http://www.tvciencia.pt/tvcnot/pagnot/tvcnot03.asp?codpub=26&codnot=8.

de Araújo, Roberto P. A., Rafael L. de Oliveira, Elder M. de Novais, Thiago D. Tadeu, Daniel B. Pereira & Ivandré Paraboni. 2010. SINotas: the Evaluation of a NLG Application. Em *Proc. Seventh International Conference on Language Resources and Evaluation*, 2388–2391.

Bateman, John & Michael Zock. 2004. Natural Language Generation. Em Ruslan Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, 284–304. Oxford University Press.

Branco, António & João Silva. 2004. Evaluating solutions for the rapid development of state-of-the-art POS taggers for portuguese. Em *4th International Conference on Language Resources and Evaluation (LREC)*, 507–510.

Denkowski, Michael & Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. Em *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, 85–91.

Denkowski, Michael & Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. Em *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 376–380.

Ferreira, Flávio, Nuno Almeida, José Casimiro Pereira, Ana Filipa Rosa, André Oliveira & António Teixeira. 2013a. Multimodal and adaptable medication assistant for the elderly. Em *8th Iberian Conference on Information Systems and Technologies*, 309–314. Lisboa.

Ferreira, Flávio, Nuno Almeida, Ana Filipa Rosa, André Oliveira, José Casimiro Pereira, Samuel Silva & António Teixeira. 2013b. Elderly centered design for interaction - the case of the S4S medication assistant. Em *Proceedings of DSAI, Procedia Computer Science*, 398–408. Vigo.

Fonseca, Ana Cristina de Sena Raposo Paiva. 1993. *Comunicação em Linguagem Natural para um Tutor Inteligente*: Universidade Técnica de Lisboa, Instituto Superior Técnico Lisboa. Tese de Mestrado.

Hall, Mark, Ian Witten & Eibe Frank. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann 3rd edn.

Hastie, Helen & Anja Belz. 2014. A comparative evaluation methodology for NLG in interactive systems. Em *9th International Conference on Language Resources and Evaluation*, 4004–4011.

Hunter, J., E. Reiter, S. G. S. (Yaji) & J. Yu. 2005. Sumtime. (Online, verificado em 23/07/2015). http://www.abdn.ac.uk/ncs/departments/computing-science/sumtime-317.php.

Hunter, James, Yvonne Freer, Albert Gatt, Ehud Reiter, Somayajulu Sripada, Cindy Sykes & Dave Westwater. 2011. BT-Nurse: computer generation of natural language shift summaries from complex heterogeneous medical data. *Journal of the American Medical Informatics Association : JAMIA* 18. 621–624.

IRSTLM. 2011. Irst language modeling toolkit. (Online, verificado em 23/07/2015). http://sourceforge.net/projects/irstlm/.

Jurafsky, Daniel & James H. Martin. 2009. *Speech and language processing*. Prentic Hall 2nd edn.

Koehn, Philipp. 2014. *Moses: Statistical machine translation system - user manual and code guide*.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin & Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. Em *45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions.*, 177–180. Praga, Rep. Checa: ACL.

Kohavi, Ron. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Em *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2* IJCAI'95, 1137–1143. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Konstantopoulos, Stasinos, Ion Androutsopoulos, Haris Baltzakis, Vangelis Karkaletsis, Colin Matheson, Athanasios Tegos & Panos Trahanias. 2008. INDIGO: Interaction with personality and dialogue enabled robots [system demonstration]. Em *18th European Conference on Artificial Intelligence*, Patras, Greece.

Langner, Brian. 2010. *Data-driven natural language generation: Making machines talk like humans using natural corpora*: School of Computer Science - Carnegie Mellon University. Tese de Doutoramento.

Langner, Brian & Alan W. Black. 2009. MOUNTAIN: A translation-based approach to natural language generation for dialog systems. Em *First International Workshop on Spoken Dialogue Systems Techology*, .

Law, Anna S., Yvonne Freer, Jim Hunter, Robert H. Logie, Neil McIntosh & John Quinn. 2005. A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. *J. Clin. Monit. Comput.* 19. 183–194.

Lemon, Oliver. 2010. Learning what to say and how to say it: joint optimization of spoken dialogue management and natural language generation. *Computer Speech & Language* 25. 210–221.

McCauley, Lee, Sidney D'Mello, Loel Kim & Melaine Polkosky. 2008. MIKI: A case study of an intelligent kiosk avatar and its usability. Em N. Magnenat-Thalmann, L. C. Jain & N. Ichalkaranje (eds.), *New Advances in Virtual Humans*, 153–176. Springer.

Mendes, Mateus Daniel. 2004. *Relações lexicais na geração de língua natural*: Universidade de Coimbra - Portugal. Tese de Mestrado.

MOSES. 2014a. Moses - baseline system. Online. http://www.statmt.org/moses/?n=Moses.Baseline.

MOSES. 2014b. Phrase-based tutorial. Online. http://www.statmt.org/moses/?n=Moses.Tutorial.

MOSES. 2014c. Syntax tutorial. Online. http://www.statmt.org/moses/?n=Moses.SyntaxTutorial.

Novais, Elder M., Rafael L. Oliveira, Daniel B. Pereira & Thiago D. Tadeu. 2009. A testbed for Portuguese Natural Language Generation. Em *Seventh Brazilian Symposium in Information and Human Language Technology*, 154–157. São Carlos, São Paulo, Brasil.

Och, Franz Josef. 2011. Giza++ statistical translation models toolkit. (Online, verificado em 23/07/2015). https://github.com/moses-smt/giza-pp.

Oliveira, Hugo Gonçalo. 2012. PoeTryMe: a versatile platform for poetry generation. Em *Proceedings of the ECAI 2012 Workshop on Computational Creativity, Concept Invention, and General Intelligence* C3GI 2012, Montpellier, France.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. Em *Meeting of the Association for Computational Linguistics*, 311–318.

Pereira, José Casimiro, António Teixeira & Joaquim Sousa Pinto. 2015. Towards a Hybrid NLG System for Data2Text in Portuguese. Em *Proceedings da 10ª Conferência Ibérica de Sistemas e Tecnologias de Informação* CISTI 2015, 679–684. Águeda, Portugal.

Portet, François, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer & Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence* 173. 789–816.

Reiter, Ehud. 2007. An architecture for data-to-text systems. Em *Proceedings of the Eleventh European Workshop on Natural Language Generation.*, 97–104. Association for Computational Linguistics.

Reiter, Ehud & Robert Dale. 1997. Building applied natural language generation systems. *Journal of Natural Language Engineering – Cambridge University Press* 3(1). 57–87.

Reiter, Ehud & Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

Ribeiro, António. 1995. *Natural Language Generation with Rhetorical Relations and Focus Theory*: Edinburgh University – UK. Tese de Mestrado.

Salzberg, Steven L. & Usama Fayyad. 1997. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery* 317–328.

Silva Junior, Douglas Fernandes Pereira, Ivandré Paraboni & Eder Miranda Novais. 2013. Um Sistema de Realização Superficial para Geração de Textos em Português. *RITA - Revista de Informática Teórica e Aplicada - Instituto de Informática da Universidade Federal do Rio Grande do Sul – Brasil* 20(3). 31–48.

Soares, Alexsandro Santos. 2001. *Gramática de Unificação Funcional: Levantamento de Requisitos para a Geração Sentencial de Português*: Instituto de Ciências Matemáticas e da Computação - USP/São Carlos - Brasil. Tese de Mestrado.

Teixeira, António, Flávio Ferreira, Nuno Almeida, Ana Filipa Rosa, José Casimiro, Samuel Silva, Alexandra Queirós & André Oliveira. 2013a. Multimodality and adaptation for an enhanced mobile medication assistant for the elderly. Em *Third Mobile Accessibility Workshop (MOBACC), CHI 2013 Extended Abstracts*, Paris.

Teixeira, António, Alexandra Queirós & Nelson Pacheco Rocha (eds.). 2013b. *Laboratório vivo de usabilidade (Living Usability Lab)*. ARC Publishing.

Turner, Ross, Somayajulu Sripada, Ehud Reiter & Ian P. Davy. 2006. Generating spatio-temporal descriptions in pollen forecasts. Em *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations.*, 163–166. Trento, Itália: Association for Computational Linguistics.

Uma Comparação Sistemática de Diferentes Abordagens para a Sumarização Automática Extrativa de Textos em Português

A Comparison of Multiple Approaches for the Extractive Summarization of Portuguese Texts

Miguel Costa e Bruno Martins
INESC-ID
Instituto Superior Técnico, Universidade de Lisboa
{miguel.angelo.costa,bruno.g.martins}@tecnico.ulisboa.pt

Resumo

A sumarização automática consiste na tarefa de gerar automaticamente versões condensadas de textos fonte, apresentando-se como um dos problemas fundamentais nas áreas da Recuperação de Informação e do Processamento de Linguagem Natural. Neste artigo, considerando metodologias puramente extrativas, são comparadas diferentes abordagens na tarefa de sumarizar documentos individuais correspondendo a textos jornalísticos escritos em Português. Através da utilização da bancada ROUGE como forma de medir a qualidade dos sumários produzidos, são reportados resultados para dois domínios experimentais diferentes, respetivamente envolvendo (i) a geração de títulos para textos jornalísticos escritos na variante Europeia do Português, e (ii) a geração de sumários com base em artigos jornalísticos escritos na variante Brasileira do Português. Os resultados obtidos demonstram que uma *baseline* simples, baseada na seleção da primeira frase, obtém melhores resultados na construção de títulos de notícias de forma extrativa, em termos de várias métricas ROUGE. No segundo domínio experimental, envolvendo a geração de sumários de notícias, o método que obteve melhores resultados foi o algoritmo LSA Squared, para as várias métricas ROUGE consideradas neste trabalho.

Palavras chave

Sumarização Automática, Avaliação Comparativa

Abstract

Automatic document summarization is the task of automatically generating condensed versions of source texts, presenting itself as one of the fundamental problems in the areas of Information Retrieval and Natural Language Processing. In this paper, different extractive approaches are compared in the task of summarizing individual documents corresponding to journalistic texts written in Portuguese. Through the use of the ROUGE package for measuring the quality of the produced summaries, we report on results for two different experimental domains, involving (i) the generation of headlines for news articles written in European Portuguese, and (ii) the generation of summaries for news articles written in Brazilian Portuguese. The results demonstrate that methods based on the selection of the first sentences have the best results when building extractive news headlines in terms of several ROUGE metrics. Regarding the generation of summaries with more than one sentence, the method that achieved the best results was the LSA Squared algorithm, for the various ROUGE metrics.

Keywords

Automatic Summarization, Comparative Evaluation

1 Introdução

A sumarização automática consiste na tarefa de gerar versões condensadas de textos fonte, apresentando-se como um dos problemas fundamentais nas áreas da Recuperação de Informação e do Processamento de Linguagem Natural (Luhn, 1958; Baxendale, 1958; Edmundson, 1969). Assim como na sumarização de textos feita por humanos, um bom sumário gerado automaticamente deve preservar as ideias principais dos textos fonte, articulando em torno destas ideias as informações centrais contidas nos documentos. Nos dias de hoje, com a crescente disponibilização de informação textual na Web, a demanda por sistemas de sumarização automática que sejam rápidos, fiáveis e robustos é maior do que nunca. Temos ainda que a produção automática de sumários tem inúmeras aplicações, sendo que estudos anteriores demonstraram já a sua utilidade a ajudar utilizadores em tarefas envolvendo a compreensão de informação em documentos textuais (i.e., sumários automáticos podem acelerar a tomada de decisões com base em informação textual (Mani et al., 2002)), ou como

forma de complementar outras aplicações e interfaces. No entanto, apesar da investigação nesta área se ter iniciado há já mais de sessenta anos, existe ainda um logo caminho a percorrer, e a tarefa está longe de estar resolvida. Argumentamos que através deste artigo outros investigadores na área podem agora ter uma ideia mais precisa de qual a performance relativa para vários dos métodos frequentemente usados na área.

Neste artigo, considerando uma metodologia extrativa para a sumarização automática de textos (i.e., os sumários são gerados pela justaposição de segmentos extraídos dos textos originais), são comparadas diferentes abordagens na tarefa de sumarizar documentos individuais correspondendo a textos jornalísticos escritos em Português. As abordagens sob comparação incluem métodos baseados em fatorização de matrizes (i.e., duas abordagens baseadas em decomposição em valores singulares, e duas abordagens baseadas em fatorização de matrizes não-negativas), métodos baseados em centralidade em grafos (i.e., adaptações do algoritmo PageRank propostas para a tarefa de sumarização automática), e *baselines* heurísticas correspondendo, por exemplo, à seleção das primeiras frases dos documentos e à seleção das frases com o maior número de palavras-chave (i.e., bi-gramas frequentes). Através da utilização da bancada ROUGE (Lin, 2004b) como forma de medir a qualidade dos sumários produzidos, são reportados resultados para dois domínios experimentais diferentes, nomeadamente:

(I) Na tarefa de gerar sumários curtos (i.e., contendo apenas uma frase) que possam ser usados como títulos de textos jornalísticos, avaliando os diferentes métodos através de uma coleção extensa de artigos escritos em Português Europeu, com cada um dos artigos associado ao título correspondente, tal como selecionado pelos editores de um portal de notícias *on-line*. Nestes testes verificou-se que o método mais simples, baseado na seleção da primeira frase de cada documento, obtém os melhores resultados.

(II) Na tarefa de gerar sumários para textos jornalísticos na variante Brasileira do Português, avaliando os diferentes métodos através de experiências com as coleções de documentos TeMário (Pardo & Rino, 2003; Maziero et al., 2007). Nestes testes verificou-se que uma abordagem baseada em decomposição de matrizes em valores singulares obtém os melhores resultados.

O restante conteúdo deste artigo encontra-se organizado da seguinte forma: A Secção 2 apresenta os principais conceitos e trabalhos anteriores na área. A Secção 3 descreve detalhadamente os vários métodos de sumarização automática que foram alvo do nosso estudo comparativo. A Secção 4 descreve a metodologia experimental considerada. A Secção 5 apresenta os resultados obtidos e uma breve análise desses resultados. Finalmente, a Secção 6 apresenta um breve resumo das principais conclusões, e descreve possíveis caminhos para trabalho futuro.

2 Trabalho Relacionado

A sumarização automática de documentos tem sido investigada ativamente na área do Processamento de Linguagem Natural (PLN) desde a segunda metade do século passado. Esta é atualmente uma área de investigação muito vasta, com inúmeros trabalhos publicados. Aos leitores deste artigo, sugere-se a consulta do trabalho de Nenkova et al. (2011) para uma visão geral sobre a área, ou a consulta do relatório técnico de Pardo (2008) para um resumo de trabalhos importantes focados no Português.

Radev et al. (2002) definiram um sumário como *um texto que é produzido a partir de um ou mais textos originais, que transmite a informação importante no(s) texto(s) original(ais), e que não é maior do que a metade do(s) texto(s) original(ais), normalmente tendo um tamanho significativamente menor.* Esta definição simples captura três aspetos importantes que caracterizam a investigação sobre o tema da sumarização automática de documentos, nomeadamente (i) os sumários podem ser produzidos a partir de um único documento ou de vários documentos, (ii) os sumários devem preservar a informação importante, e (iii) os sumários devem ser curtos. Dado que o fluxo e a densidade de informação, num determinado documento, é geralmente não-uniforme (i.e., algumas partes dos documentos são mais importantes do que outras), o grande desafio que se apresenta à sumarização automática consiste em discriminar as partes mais informativas de um documento.

O texto introdutório de Radev et al. (2002) também apresenta alguma terminologia importante na área da sumarização automática. Temos assim que *extração* se refere ao processo de identificação de segmentos importantes de um texto original, reproduzindo-os na íntegra aquando da geração de sumários. Por outro lado, *abstração* é um processo que tem como objetivo produzir conteúdos textuais novos que reflitam

os aspetos importantes de uma dada fonte textual. Um processo de *fusão* combina segmentos extraídos de forma coerente. Finalmente, um processo de *compressão* visa remover segmentos pouco importantes dos textos produzidos como sumários (Coster & Kauchak, 2011). Importa notar que, enquanto a sumarização automática extrativa se preocupa principalmente com o conteúdo dos sumários, baseando-se geralmente apenas na extração de frases, a sumarização automática abstrativa coloca uma forte ênfase na forma, com o objetivo de produzir sumários gramaticalmente corretos, o que geralmente requer técnicas avançadas de geração de linguagem natural (i.e., a sumarização abstrativa normalmente envolve a fusão da informação extraída, a compressão de frases, e a reformulação de frases (Knight & Marcu, 2002; Jing & McKeown, 2000; Almeida & Martins, 2013)). Embora um sumário abstrativo possa ser mais conciso, os sumários puramente extrativos são mais viáveis computacionalmente, e estas abordagens mais simples tornaram-se o padrão no campo da sumarização automática de textos.

Os primeiros trabalhos de investigação, abordando a sumarização automática extrativa de documentos textuais escritos em Inglês, propuseram métodos heurísticos para extrair as frases mais importantes de documentos individuais, usando combinações de atributos como a frequência de palavras ou de segmentos de texto (Luhn, 1958), a posição no texto (Baxendale, 1958), ou a ocorrência de segmentos-chave inferidos para os textos (Edmundson, 1969). Por exemplo na abordagem proposta por Luhn (1958), as palavras são inicialmente transformadas nos seus radicais (i.e., é efetuado um processo de *stemming*), e as *stop-words* comuns são removidas. Luhn compilou uma lista de palavras significativas (i.e., palavras associadas a conteúdos semânticos importantes), classificando-as por ordem decrescente de frequência, sendo que o índice de cada palavra nesta lista ordenada proporciona uma medida da sua importância. Ao nível de cada frase, um fator de importância é derivado das palavras que a constituem, refletindo o número de ocorrências de palavras significativas dentro da frase, e a distância linear entre elas (i.e., considerando a possível utilização de palavras não significativas nas frases, entre as ocorrências de palavras significativas). As várias frases de um documento são classificados por ordem do seu fator de importância, e as primeiras frases nesta ordenação são finalmente selecionadas para formar o sumário. Vários trabalhos posteriores consideraram ideias semelhantes às propostas nestes trabalhos seminais, concentrando-se na aplicação a outros idiomas ou a domínios de aplicação específicos, com especial foco no caso de textos jornalísticos. No caso particular da língua Portuguesa, a grande maioria dos trabalhos anteriores estudou a aplicação de abordagens heurísticas semelhantes às descritas atrás.

Na década de 1990, com a crescente popularização do uso de técnicas de aprendizagem automática em tarefas de PLN, surgiram uma série de publicações envolvendo abordagens estatísticas para a produção automática de sumários. Por exemplo Kupiec et al. (1995) descrevem um método derivado do trabalho de Edmundson (1969), que era capaz de aprender a partir de dados anotados manualmente. Uma função de classificação, baseada na abordagem naïve-Bayes, era usada para categorizar cada frase como interessante de considerar (i.e., de extrair) para o sumário ou não. As características usadas pela função de classificação eram muito semelhantes às do trabalho de Edmundson (1969), mas estes autores incluíram ainda o comprimento da frase e a presença de palavras em maiúsculas. A cada frase era atribuída uma pontuação de acordo com a probabilidade obtida pelo classificador naïve-Bayes, e as n frases melhor pontuadas formam o sumário.

Aone et al. (1999) também utilizaram um classificador naïve-Bayes, mas neste caso com um conjunto de características mais ricas (e.g., considerando características tais como pesos obtidos pela heurística *term-frequency* $\times$ *inverse document frequency* (TF-IDF) para cada um dos termos presentes nos documentos, como forma de tentar capturar conceitos-chave nos documentos a sumarizar). Neste trabalho, além de palavras individuais, os autores consideraram ainda o uso de colocações relevantes (i.e., bi-gramas de substantivos) calculadas estatisticamente, ou o uso de entidades mencionadas nos textos, como unidades de contagem. Os autores também utilizaram técnicas simples como forma de resolver alguns tipos de co-referências nos textos (e.g., associar os acrónimos dentro de um documento a um mesmo conceito único, como *EUA* a *Estados Unidos* ou como *IBM* a *International Business Machines*, por forma a aumentar a coesão nas representações). Sinónimos e variantes morfológicas também foram fundidas ao considerar os termos lexicais individuais, sendo os mesmos identificados através da WordNet (Miller, 1995).

Lin (1999) rompeu com o pressuposto de modelação em que as características usadas na classificação são independentes umas das outras, tentando modelar o problema da extração de frases importantes usando árvores de decisão e com um

amplo conjunto de características, em vez de usar um classificador naïve-Bayes. Osborne (2002), por sua vez, utilizou modelos de máxima entropia treinados por um método de gradiente descendente conjugado, considerando características como pares de palavras, com todas as palavras truncadas para um máximo de dez carateres, o comprimento das frases, a posição das frases no texto, e outras características simples tais como a ocorrência das frases dentro de secções de introdução ou de conclusão.

Conroy & O'Leary (2001) abordaram o problema da extração de frases desde documentos de texto através de modelos de Markov com variáveis ocultas (HMMs), numa tentativa de capturar dependências locais entre frases. Apenas três características representativas das frases foram utilizadas neste trabalho, nomeadamente a posição da frase no documento (i.e., informação embutida na própria estrutura de estados do HMM), o número de termos na frase, e a probabilidade de observar os termos da frase, dados os termos do documento. Além de HMMs, outros trabalhos anteriores basearam-se em modelos sequenciais mais sofisticados, por exemplo considerando o formalismo dos *Conditional Random Fields* (CRFs) e utilizando conjuntos de características mais ricos (Shen et al., 2007).

As propostas baseadas em aprendizagem automática apresentam geralmente a desvantagem de necessitarem de dados de treino sob a forma de frases extraídas desde documentos de texto (i.e., os exemplos de treino consistem de frases anotadas como interessantes ou não de pertencer a um sumário extrativo). Explorando a convenção de que as partes mais importantes de um texto jornalístico são geralmente colocadas nos parágrafos iniciais (i.e., são poucas as técnicas de sumarização automática extrativa que conseguem resultados significativamente melhores do que uma abordagem simplista baseada em extrair as primeiras frases), Svore et al. (2007) propuseram e avaliaram uma abordagem baseada em redes neuronais (i.e., baseada num algoritmo de *learning to rank* denominado RankNet), na qual se treina um modelo a partir de características (e.g., frequências de n-gramas) extraídas desde frases em textos jornalísticos e das suas respetivas posições nos textos. O modelo aprende a inferir qual a posição que seria mais adequada para cada frase, sendo que posteriormente as frases associadas às primeiras posições são as selecionadas para a elaboração do sumário. Os autores utilizaram também, nos seus modelos, características derivadas de recursos de informação externos, tais como artigos da Wikipédia ou históricos de pesquisas efetuadas no motor de busca de notícias da Microsoft, sob a conjetura de que frases de um documento que contenham palavras frequentemente usadas nas pesquisas do motor de busca, ou que contenham entidades correspondentes a artigos da Wikipédia, devem ser consideradas como interessantes para colocar nos sumários.

Como forma de contornar a necessidade de obtenção de dados de treino, vários trabalhos anteriores exploraram ainda técnicas não-supervisionadas, por exemplo baseadas em fatorização de matrizes. Intuitivamente, estes métodos tentam agrupar as frases de um documento em grupos/componentes que sejam coerentes entre si, escolhendo posteriormente as frases mais representativas de cada grupo, por forma a construir os sumários. Por exemplo Gong & Liu (2001) propuseram um método que utiliza análise semântica latente (LSA) para selecionar frases adequadas à construção de um sumário. Este método cria inicialmente uma matriz de *termos* × *frases*, onde cada coluna representa o vetor de frequências ponderadas dos termos de uma frase. De seguida, é aplicada uma decomposição em valores singulares (SVD) sob a matriz, por forma a derivar a estrutura semântica latente. A(s) frase(s) com maior(es) peso(s) no primeiro conceito latente (i.e., o conceito correspondente ao primeiro valor singular) são finalmente selecionadas para a formação do sumário (i.e., o método escolhe a(s) frase(s) mais informativa(s) do primeiro valor singular). Steinberger & Ježek (2004) propuseram uma abordagem alternativa também baseada na decomposição SVD, capaz de produzir resultados de melhor qualidade, onde se usam os vários valores singulares. Por outro lado, autores como Lee et al. (2009) ou como Mashechkin et al. (2011) propuseram a utilização de fatorização de matrizes não-negativas (NMF), decompondo a matriz de *termos* × *frases* em factores não negativos, por forma a extrair as frases com maior pontuação em cada um dos componentes latentes descobertos desta forma.

Outros autores ainda propuseram a utilização de métodos não-supervisionados baseados em grafos, como forma de capturar as frases mais centrais para a construção de sumários extrativos. Os métodos baseados em grafos começam normalmente pela construção de um grafo que represente o documento, ou coleções de documentos, a sumarizar. Nesta representação, cada nó do grafo é geralmente uma frase e, se a similaridade entre um par de frases estiver acima de um dado valor limiar, então existe uma aresta entre o par de frases. As frases centrais são selecionadas para formar os sumários através de um

processo de votação pelas suas frases vizinhas. Por exemplo Erkan & Radev (2004) propuseram um algoritmo chamado LexRank para calcular a importância de cada frase, com base no conceito de *eigenvector centrality* (i.e., uma noção de prestígio dos nós, semelhante à que se encontra associada ao algoritmo PageRank do Google (Page et al., 1999; Franceschet, 2011)). Outros métodos semelhantes foram propostos por Mihalcea (2004), por Mihalcea & Tarau (2005), ou por Wan & Yang (2008). Yeh et al. (2005) propuseram um método que combina as ideias de análise semântica latente (LSA) e centralidade em grafos. As representações semânticas das frases, obtidas por decomposição em valores singulares, são usadas para construir um grafo de relações entre as frases. Finalmente, é aplicada uma medida de significância dos nós do grafo, baseada no trabalho original de Salton et al. (1997), e são escolhidas as k frases mais conectadas no grafo, sendo as mesmas apresentadas de acordo com a ordem com que as frases surgem no documento original.

Alguns esforços anteriores focaram-se, por sua vez, no agregar dos resultados de vários métodos diferentes de sumarização automática. Thapar et al. (2006) apresentaram uma abordagem de meta-sumarização baseada em grafos, que compara grafos gerados com base em cada um dos sumários produzidos pelos métodos individuais, com um grafo que agrega os resultados dos diferentes métodos de sumarização. Wang & Li (2010) avaliaram sistematicamente diferentes métodos para a combinação dos resultados de sistemas de sumarização extrativa (i.e., diferentes esquemas de agregação de ordenações de frases), propondo depois um método de consenso ponderado para agregar os resultados de vários métodos.

Vários trabalhos anteriores abandonaram a sumarização automática de documentos individuais, em vez disso considerando múltiplas fontes de informação (i.e., sumarização multi-documento) que se podem sobrepor e complementar, ocasionalmente apresentando ainda contradições. Na sumarização multi-documento, as principais tarefas relacionam-se não só com identificar e lidar com redundância em documentos, mas também com o reconhecer conteúdos novos e com o garantir que o sumário final é coerente e completo. Técnicas extrativas foram aplicadas à sumarização automática multi-documento, fazendo por exemplo uso de medidas de similaridade entre pares de frases. As abordagens propostas variam essencialmente na forma como estas semelhanças são utilizadas: alguns trabalhos procuram identificar temas comuns através do agrupamento (i.e., *clustering*) de frases e, em seguida, selecionam uma frase para representar cada grupo (McKeown & Radev, 1995; Radev et al., 2000), enquanto outros métodos geram uma frase composta de elementos extraídos de cada grupo (Barzilay et al., 1999), e outros autores ainda estudaram abordagens dinâmicas que incluem cada passagem candidata apenas se a mesma for considerada nova no que diz respeito às passagens incluídas anteriormente, através do conceito da relevância marginal máxima (Carbonell & Goldstein, 1998). Alguns autores abordaram ainda a sumarização multi-documento em conjunto com a compressão de frases (Almeida & Martins, 2013), enquanto outros autores estudaram o problema da sumarização multi-documento em diferentes variantes multilingues (Litvak et al., 2010; Siddharthan & McKeown, 2005; Fung & Ngai, 2006).

Importa ainda referir que alguns métodos modernos para a sumarização automática, em particular no caso da sumarização multi-documento ou no caso de métodos que combinam a compressão de frases com a extração de frases relevantes, se baseiam no formalismo da programação linear inteira (ILP). Nestas abordagens, a tarefa de sumarização é vista como um problema de otimização combinatória, em que se pretende selecionar um conjunto de frases, até um tamanho máximo preestabelecido, que maximize a soma das pontuações de relevância, e que ao mesmo tempo minimize a redundância entre as frases selecionadas, de um conjunto de frases obtido através do processamento do(s) documento(s) e considerando diferentes taxas de compressão para as frases originais (Hirao et al., 2009; Almeida & Martins, 2013). No entanto, os métodos de sumarização considerados no nosso estudo comparativo baseiam-se apenas na seleção das frases melhor pontuadas em termos da sua relevância, não tendo sido considerada a combinação de pontuações de relevância com outras medidas (e.g., capturando a redundância entre as várias frases seleccionadas).

3 Métodos de Sumarização Automática Considerados no Estudo Comparativo

Neste artigo, foram realizadas experiências comparativas envolvendo diferentes métodos de sumarização automática extrativa. As abordagens sob comparação incluem métodos baseados em fatorização de matrizes (i.e., duas abordagens baseadas em decomposição em valores singulares, e duas baseadas em fatorização de matrizes não-negativas), métodos baseados em grafos

(i.e., adaptações do algoritmo PageRank propostas para a tarefa de sumarização automática), e *baselines* heurísticas, correspondendo à seleção das primeiras frases dos documentos, e à seleção das frases com maior relevância em termos da ocorrência de termos-chave, especificamente considerando bi-gramas frequentes.

Os vários métodos de sumarização automática em estudo partilham uma fase de pré-processamento comum, em que os textos a sumarizar são segmentados em palavras e em frases, através do uso dos mecanismos de segmentação de textos, comuns a várias línguas Indo-Europeias, disponíveis num pacote Python para PLN denominado nltk[1]. Em todos os métodos testados, temos que as diferentes frases dos documentos foram representadas através de conjuntos de palavras-chave, tendo sido removidas *stop-words* comuns, e tendo os textos sido convertidos por forma a se utilizarem apenas carateres minúsculos e sem acentos.

As seguintes subsecções apresentam em detalhe cada um dos métodos considerados, os quais permitem essencialmente pontuar as frases de um dado texto fonte, de acordo com a sua adequabilidade para pertencerem a um sumário. A produção dos sumários com base nos métodos descritos de seguida baseia-se na seleção da(s) frase(s) melhor pontuada(s), concatenando-as pela ordem na qual surgem no texto original. As abordagens de sumarização consideradas neste artigo são desta forma puramente extrativas, muito embora para trabalho futuro se considere também a integração de abordagens de compressão de frases (Yamangil & Nelken, 2008; Coster & Kauchak, 2011; Bach et al., 2011).

3.1 Decomposição em Valores Singulares

A análise semântica latente (LSA) é um método algébrico frequentemente utilizado na área do PLN para analisar as relações entre documentos e os termos neles contidos, através da produção de um conjunto de conceitos latentes relacionados com os documentos e os termos. Este método assume que as palavras semanticamente relacionadas tendem a coocorrer nos textos. No contexto da aplicação em sumarização automática, uma matriz esparsa A representando as ocorrências de termos por frases (i.e., uma matriz com m linhas que representam os termos únicos, e com n colunas que representam cada frase) é construída a partir de um documento original. De seguida, é calculada uma decomposição em valores singulares (SVD), tipicamente com o objetivo de reduzir o número de linhas (i.e., os termos correlacionados são agrupados em conceitos, capturando assim fenómenos como a sinonímia entre termos), preservando a estrutura de similaridade entre as colunas (i.e., entre as frases). A decomposição SVD é tipicamente calculada com base num algoritmo que envolve duas etapas, o qual começa por reduzir a matriz A à sua forma bi-diagonal (i.e., a uma matriz com entradas diferentes de zero apenas na diagonal principal e nos valores que se encontram acima ou abaixo), e que de seguida calcula a decomposição SVD da matriz bi-diagonal através de um método iterativo.

Resumidamente, temos que a decomposição SVD fatoriza a matriz A em três matrizes, U, D e V^T, de tal forma que $A = UDV^T$. A matriz $U = [u_{ij}]$, de dimensão $m \times n$, é uma matriz unitária cujas colunas são vetores ortonormais, denominados os vetores singulares à esquerda. Por sua vez $D = \text{diag}(\alpha_1, \alpha_2, \ldots, \alpha_n)$ é uma matriz diagonal $n \times n$, cujos elementos diagonais são valores singulares não negativos, ordenados de forma decrescente. Finalmente, temos que $V^T = [v_{ji}]$ é uma matriz ortogonal de dimensão $n \times n$, cujas colunas são chamadas os vetores singulares à direita.

No contexto da aplicação em sumarização automática, um algoritmo simples pode ser usado para selecionar a(s) melhor(es) frase(s), com base na decomposição SVD e usando a matriz de valores singulares à direita V^T. Cada frase i é representada pelo vetor coluna $\psi_j = [v_{j1}v_{j2}, \ldots, v_{jn}]^T$ da matriz V^T e, com uma abordagem simples, basta-nos selecionar o primeiro vetor singular direito da matriz V^T e, em seguida, a(s) frase(s) que têm o maior valor índice no vetor. Seguidamente, se necessário, efetua-se o mesmo processo para o segundo vetor singular direito da matriz, até se chegar ao número desejado de frases selecionadas para a construção do sumário (Gong & Liu, 2001). Neste trabalho denominamos esta abordagem de LSA Classic.

Uma outra abordagem para selecionar a(s) melhor(es) frase(s) a partir da SVD foi proposta por Steinberger & Ježek (2004). Em vez de selecionar a(s) frase(s) de topo do primeiro vetor singular, a(s) frases são selecionadas com base num peso global obtido de todos os vetores singulares. Para cada vetor coluna da matriz V^T (i.e., para cada frase j), vamos calcular a raiz quadrada do quadrado dos seus componentes multiplicados pelo quadrado dos valores singulares correspondentes na matriz D, desta forma favorecendo os valores do índice na matriz V^T que correspondem aos maiores valores singulares. Em seguida, escolhem-se as frases com o maior peso com-

[1] http://www.nltk.org/

binado em todos os componentes importantes. De uma forma resumida, temos que de acordo com este método escolhemos a(s) frase(s) j que tenha(m) o(s) valor(es) de relevância mais elevado(s), tal como produzidos por:

$$s_j = \sqrt{\sum_{i=1}^{n} v_{j,i}^2 \times \alpha_i^2} \quad (1)$$

Na equação, s_j é o peso do vetor de termos j no modificado espaço de vetores latente e n é o número de dimensões do novo espaço. Denominamos esta abordagem por LSA Squared.

Nos testes aqui reportados, para efetuar a fatorização SVD das matrizes de termos por frases, foi utilizada a implementação de um pacote Python denominado scikit-learn[2].

3.2 Fatorização Não Negativa

A fatorização de matrizes não-negativas (NMF) é um método de decomposição de matrizes recentemente desenvolvido, o qual impõe a restrição de que as entradas e os factores resultantes devem ser não-negativos, ou seja, todos os elementos das matrizes resultantes da decomposição têm de ser iguais ou maiores que zero. Na proposta original de Lee & Seung (1999) a fatorização em matrizes não-negativas decompõe uma matriz A com m linhas e n colunas, correspondendo à representação das n frases contendo m termos, em duas matrizes não-negativas W e H, de forma a que $A_{m\times n} \approx W_{m\times r} \times H_{r\times n}$, onde $W_{m\times r}$ é uma matriz não-negativa de características semânticas (i.e., a matriz dos termos) e onde $H_{r\times n}$ é uma matriz não negativa de variáveis semânticas (i.e., a matriz das frases). Um dos algoritmos mais populares para encontrar decomposições NMF é baseado numa regra de atualização multiplicativa, que atualiza iterativamente as matrizes W e H até obter a convergência de uma função objetivo do tipo $J = ||A - WH||^2$ sob um determinado valor limiar predefinido, ou até o algoritmo exceder um determinado número de passos. As seguintes regras de atualização são utilizadas em cada passo do algoritmo iterativo:

$$H_{\alpha,\mu} \leftarrow H_{\alpha,\mu} \times \frac{(W^T A)_{\alpha,\mu}}{(W^T W H)_{\alpha,\mu}} \quad (2)$$

$$W_{i,\alpha} \leftarrow W_{i,\alpha} \times \frac{(AH^T)_{i,\alpha}}{(WHH^T)_{i,\alpha}} \quad (3)$$

Depois de encontrar a decomposição NMF, podemos calcular uma medida genérica de relevância para cada frase, de acordo com a proposta original de Lee et al. (2009), a qual corresponde à seguinte equação:

$$\mathrm{GR}_j = \sum_{i=1}^{r} \left(H_{i,j} \times \frac{\sum_{q=1}^{n} H_{i,q}}{\sum_{p=1}^{r}\sum_{q=1}^{n} H_{p,q}} \right) \quad (4)$$

A(s) frase(s) com valor(es) mais elevado(s) em termos da medida de relevância são finalmente selecionadas para a formação do sumário. Denominamos esta abordagem, neste trabalho, pela sigla NMF GR.

Outra abordagem baseada em NMF corresponde à proposta de Mashechkin et al. (2011), onde também se pretende encontrar uma medida de relevância para cada frase, mas de uma forma mais abrangente, utilizando a seguinte equação:

$$\mathrm{ExtR}_j = \sum_{i=1}^{k} \left(\|W_i\|^2 \times \|H_i\| \times H_{i,j} \right) \quad (5)$$

Na equação $\|W_i\|^2$ é o quadrado da norma Euclideana do vetor W_i e $\|H_i\|$ é a norma Euclideana dos tópicos do vetor H_i. A(s) frase(s) com valor(es) mais elevado(s) em termos desta medida de relevância são finalmente selecionadas para a formação do sumário. Denominamos esta abordagem por NMF ExtR.

Nos testes aqui reportados, para efetuar a fatorização NMF das matrizes, foi utilizada a implementação do pacote scikit-learn.

3.3 Centralidade em Grafos

Além de abordagens baseadas em fatorização de matrizes, foram também feitos testes com métodos baseados em grafos para a sumarização automática de textos, seguindo as ideias originalmente apresentadas por Mihalcea (2004). Para efetuar os testes em que se usam algoritmos baseados em grafos foi utilizado um pacote Python denominado Networkx[3]. Por forma a suportar a aplicação de algoritmos de ordenação baseados em grafos, em tarefas de sumarização automática, começamos por construir um grafo que represente o documento a sumarizar, interligando as suas frases através de relações de similaridade que capturem a sobreposição de conteúdos textuais. Estas relações entre pares de frases podem ser vistas como um processo de recomendação, em que uma frase que aborda alguns conceitos de um texto dá ao leitor uma recomendação, no sentido de ele se poder referir a outras frases no

[2] http://scikit-learn.org/

[3] http://networkx.github.io/

texto que abordem os mesmos conceitos (Mihalcea, 2004). Nas nossas experiências, foram testadas as seguintes métricas de similaridade diferentes, aplicando posteriormente uma variante do algoritmo PageRank (Page et al., 1999), para grafos pesados e não-direcionados:

- O coeficiente de similaridade de Jaccard entre as frases, com base em termos individuais. Este coeficiente mede a similaridade entre dois conjuntos de termos A e B, sendo definido como o tamanho da intersecção dos conjuntos dividido pelo tamanho da sua união:

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (6)$$

- A similaridade do cosseno entre as frases, com base em termos individuais. São calculadas as mesmas medidas de pesagem, utilizadas nos testes envolvendo decomposição de matrizes, para os termos em cada uma das frases, o que dá origem a uma matriz de *termos* × *frases*. Seguidamente é calculado o cosseno do ângulo θ formado entre os vetores que representem pares de frases A e B, de acordo com a seguinte equação:

$$\text{sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (7)$$

Em cada uma das abordagens anteriores, o cálculo da similaridade entre pares de frases resulta num grafo denso (i.e., todos os nós encontram-se interligados entre si), em que existem pesos associados a cada aresta não-direcionada, indicando a força das associações entre os vários pares de frases no texto. Sob este grafo, é executada uma versão ligeiramente modificada do algoritmo PageRank (PR) (Page et al., 1999), por forma a obter uma ordenação dos nós do grafo de acordo com o seu prestígio/importância. Esta versão adaptada permite contabilizar pesos tal como associados às arestas, correspondendo à seguinte equação, onde $\text{Ln}(V_i)$ corresponde ao conjunto de nós ligados no grafo ao nó V_i, e onde $w_{j,i}$ corresponde ao peso da aresta que liga os nós j e i.

$$\text{PR}(V_i) = \frac{(1-d)}{N} + d \times \sum_{V_j \in \text{Ln(V}_\text{i})} \frac{w_{j,i}}{\sum_{V_k \in \text{Ln}(V_j)} w_{j,k}} \text{PR}(V_j) \quad (8)$$

Da equação acima, é possível verificar que cada nó V_i do grafo terá uma pontuação que depende de um fator inicial $(1-d)$, uniformemente associado cada um dos N nós e que normalmente é escolhido com base no valor $d = 0.85$. A pontuação de cada nó depende também das pontuações associadas aos nós que lhe estão diretamente associados. O cálculo do PageRank, normalmente efetuado de forma iterativa, produz, como resultado, uma distribuição de probabilidade sobre os nós do grafo, em que cada nó terá uma probabilidade correspondente à sua importância, tal como derivada das associações para com todos os restantes nós do grafo. Depois do algoritmo PageRank ser executado sob o grafo, as frases (i.e., os nós do grafo) são ordenadas pela sua pontuação, e a(s) frase(s) de topo são selecionadas para inclusão no sumário, do documento. No nosso estudo denominamos esta abordagem por TextRank.

Foi também testado o uso de uma probabilidade inicial não-uniforme para cada frase. O grafo é construído de forma semelhante ao caso do TextRank, ou seja, é criado um grafo não-direcionado, interligando cada frase do documento a sumarizar através de relações de similaridade. Uma distribuição de probabilidades inicial, sob cada um dos N nós, é calculada através da seguinte fórmula, a qual substitui a primeira parte da Equação 8:

$$\text{PR}(V_i) = \frac{\frac{1}{\text{POS}(V_i)^\alpha}}{\sum_{j=1}^{N} \frac{1}{\text{POS}(V_j)^\alpha}} \times (1-d) + d \times \ldots \quad (9)$$

Nesta equação o parâmetro α foi usado com o valor de 0.85, tendo este valor sido selecionado por ter obtido os melhores resultados num conjunto inicial de experiências. A função $\text{POS}(V_i)$ refere-se a posição de cada frase V_i no documento, dando-se desta forma um peso superior às primeiras frases. Neste trabalho denominamos esta abordagem por TextRank Init.

3.4 Abordagens Heurísticas

Foram ainda efetuados testes com dois métodos *baseline* inspirados em heurísticas simples que foram propostas em alguns dos trabalhos seminais na área da sumarização automática.

O primeiro destes métodos *baseline* pretende demonstrar a convenção de que as partes mais importantes de um texto jornalístico são geralmente colocadas nos parágrafos iniciais, tendo para isso sido comparadas duas variantes distintas desta ideia. A primeira variante consiste em selecionar a(s) primeira(s) frases de cada documento, e a segunda variante consiste em selecionar a(s) frase(s) de forma aleatória. O segundo método *baseline* procura selecionar a(s) frase(s) onde ocorrem mais vezes os conceitos chave do documento. São primeiro extraídos os n conceitos chave (i.e., os n bi-gramas de palavras com maior número de ocorrências) mais importantes do documento e, para cada frase, são somadas as pontuações associadas aos conceitos chave que nelas ocorram. Por fim, são selecionadas as frases com maior pontuação agregada.

4 Avaliação Experimental

A avaliação comparativa dos métodos de sumarização automática, descritos na secção anterior, foi feita com base em dois domínios experimentais diferentes, envolvendo (i) a geração de títulos para textos jornalísticos escritos na variante Europeia do Português, e (ii) a geração de sumários com base em artigos jornalísticos escritos na variante Brasileira do Português.

No primeiro caso, foram usados 40.000 textos jornalísticos publicados originalmente no portal *sapo.pt*, associados aos respetivos títulos tal como selecionados pelos editores do portal. A tarefa a resolver consiste em gerar sumários curtos (i.e., de uma frase apenas) com base no texto das notícias, que se apresentem como bons títulos para os artigos (i.e., que sejam muito semelhantes aos títulos escolhidos pelos editores). Para a criação do corpus do *sapo.pt* apenas foram selecionadas notícias que tivessem no mínimo sete frases, número selecionado após o cálculo da média do número de frases nos textos de todo o corpus a que tivemos acesso. Após a seleção das notícias, o corpus do *sapo.pt* contém em média doze frases por documento e uma média de cento e quarenta e quatro palavras por notícia.

No segundo caso, foram usados os textos jornalísticos associados ao TeMário (sigla para *TExtos com suMÁRIOs*), um conjunto de dados criado originalmente em 2003 e posteriormente revisto em 2006, sendo que a tarefa a resolver consiste em gerar sumários para artigos jornalísticos individuais, semelhantes aos sumários criados por peritos humanos. O TeMário 2006 é um corpus de 150 textos jornalísticos na variante Brasileira do Português associados aos seus respetivos sumários (Maziero et al., 2007), construído para complementar o corpus TeMário original (Pardo & Rino, 2003), o qual contém 100 textos e sumários da mesma natureza. Ambos os corpora tiveram os seus resumos produzidos por especialistas humanos. As várias notícias associadas aos textos dos corpora TeMário contêm, em média, dezasseis frases e uma média de cento e trinta e cinco palavras por notícia, após a remoção de *stop-words*. Desta forma, no caso dos testes com estes dados, cada um dos métodos em estudo foi usado para selecionar trinta por cento do número de frases dos textos originais, para a construção dos sumários.

Em termos das métricas consideradas para a avaliação, temos que Lin (2004b) introduziu um conjunto de métricas denominadas *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE), que deste então se tornaram a norma em termos de métricas para a avaliação automática de sistemas de sumarização. Estas métricas encontram-se implementadas num pacote de software[4] disponível livremente para a avaliação de sistemas de sumarização, o qual foi usado no contexto das nossas experiências.

Consideremos um conjunto de sumários de referência $R = \{r_1, \ldots, r_m\}$, e um sumário s gerado automaticamente. Consideremos ainda um vetor binário $\Phi_n(d)$ que representa os n-gramas contidos num documento d, onde o i-ésimo componente ϕ_n^i tem o valor 1 se o i-ésimo n-grama se encontra contido em d, e 0 caso contrário. A métrica ROUGE-N apresenta-se como uma estatística que captura a cobertura dos n-gramas, sendo calculada da seguinte forma:

$$\text{ROUGE-N}(s) = \frac{\sum_{r \in R} \langle \Phi_n(d), \Phi_n(s) \rangle}{\sum_{r \in R} \langle \Phi_n(d), \Phi_n(r) \rangle} \quad (10)$$

Na fórmula acima, $\langle ., . \rangle$ representa a definição usual para o produto interno de vetores. A métrica ROUGE-N, tal como definida acima, pode ser usada em cenários de avaliação em que existam múltiplos sumários de referência, muito embora a mesma também possa tomar apenas o sumário de referência mais similar para com o sumário gerado automaticamente:

$$\text{ROUGE-N}_{multi}(s) = \max_{r \in R} \frac{\langle \Phi_n(d), \Phi_n(s) \rangle}{\langle \Phi_n(d), \Phi_n(r) \rangle} \quad (11)$$

Uma outra métrica proposta por Lin (2004b) aplica o conceito da sub-sequência comum mais longa (LCS). O racional por detrás desta ideia prende-se com o facto de que quanto maior for a

[4] http://www.berouge.com/

LCS entre duas frases de sumários, maior a similaridade entre elas. Consideremos um conjunto de frases de referência $r_1, \ldots, r_u$ para os documentos em R, e consideremos um sumário candidato s (i.e., s corresponde à concatenação de frases extraídas por um sumarizador extrativo). A métrica ROUGE-L é definida como uma média harmónica calculada com base na LCS, tal como apresentada na seguinte equação:

$$\text{ROUGE-L}(s) = \frac{(1+\beta^2) \times \text{R}_{\text{LCS}}(s) \times \text{P}_{\text{LCS}}(s)}{\text{R}_{\text{LCS}}(s) + \beta^2 \times \text{P}_{\text{LCS}}(s)} \quad (12)$$

Na fórmula acima,

$$\text{R}_{\text{LCS}}(s) = \frac{\sum_{i=1}^{u} \text{LCS}(r_i, s)}{\sum_{i=1}^{u} |r_i|}$$

e

$$\text{P}_{\text{LCS}}(s) = \frac{\sum_{i=1}^{u} \text{LCS}(r_i, s)}{|s|},$$

sendo que $|x|$ denota o tamanho de uma frase x, e $\text{LCS}(x, y)$ denota o tamanho da sub-sequência comum mais longa entre as frases x e y. O parâmetro real β controla o balanceamento entre os componentes $\text{R}_{\text{LCS}}(s)$ e $\text{P}_{\text{LCS}}(s)$, tomando normalmente o valor de 1 (neste caso, a medida ROUGE-L corresponde exatamente a uma média harmónica). A função $\text{LCS}(x, y)$ pode ser calculada através de uma abordagem de programação dinâmica.

Uma outra métrica também introduzida por Lin (2004b) é a ROUGE-S, a qual pode ser vista como uma versão com intervalos da métrica ROUGE-N, com $N = 2$ (i.e., uma métrica baseada em *skip bi-grams*). Consideremos um vetor binário $\Psi_2(d)$ indexado por pares ordenados de palavras, onde o componente $\psi_2^i(d)$ toma o valor 1 caso o i-ésimo par seja uma sub-sequência de palavras existente em d, e 0 caso contrário. A métrica ROUGE-S pode ser calculada como:

$$\text{ROUGE-S}(s) = \frac{(1+\beta^2) \times \text{R}_{\text{S}}(s) \times \text{P}_{\text{S}}(s)}{\text{R}_{\text{S}}(s) + \beta^2 \times \text{P}_{\text{S}}(s)} \quad (13)$$

Na fórmula acima, temos que o parâmetro

$$\text{R}_{\text{S}}(s) = \frac{\sum_{i=1}^{u} \langle \Psi_2(r_i), \Psi_2(s) \rangle}{\sum_{i=1}^{u} \langle \Psi_2(r_i), \Psi_2(r_i) \rangle},$$

enquanto que o parâmetro

$$\text{P}_{\text{S}}(s) = \frac{\sum_{i=1}^{u} \langle \Psi_2(r_i), \Psi_2(s) \rangle}{\langle \Psi_2(s), \Psi_2(s) \rangle}.$$

Nos nossos testes, foi usada uma versão estendida do ROUGE-S denominada ROUGE-SU, com o número de *skip bi-grams* = 4 e que considera também a sequência de uni-gramas. A métrica ROUGE-SU pode ser obtida através da métrica ROUGE-S, ao adicionar marcadores de início e fim nas frases candidatas e de referência.

As várias versões da medida ROUGE foram avaliadas no passado, medindo a correlação para com avaliações produzidas por peritos humanos (Lin, 2004a,b). A variante ROUGE-2 apresenta-se como a melhor de entre as várias variantes da ROUGE-N, e as medidas ROUGE-L e ROUGE-SU todas apresentaram bons resultados. Para as experiências efetuadas no contexto deste artigo, são apresentados resultados sob todas estas métricas, considerando $N = 1, 2$, para o caso da métrica ROUGE-N por serem estes os valores mais usados na área, e considerando *skip bi-grams* de tamanho 4, para o caso da métrica ROUGE-SU.

Para as abordagens baseadas em decomposição de matrizes ou baseadas em grafos, testadas neste estudo, compararam-se diferentes métodos de pesagem dos termos contidos nos documentos, tal como especificados na Tabela 1.

Quanto às abordagens baseadas no algoritmo TextRank, foram testados diversos valores relativamente ao parâmetro correspondente ao número de iterações (i.e., 100, 200, 300, 400, 500). No entanto, não se obtiveram alterações significativas nos resultados, tendo sido selecionado o número mínimo de iterações testado (i.e., 100) para a apresentação dos resultados neste estudo. Nas Figuras 1 e 2 apresentamos os resultados para cada um dos métodos baseados em grafos que foram considerados, para ambos os corpora.

Para o caso dos métodos baseados nas ocorrências de bi-gramas frequentes, testaram-se diferentes números de bi-gramas k para a construção da lista de k bi-gramas mais importantes, aquando da geração dos sumários. Os resultados são apresentados na Figura 3.

No caso dos métodos baseados na decomposição SVD, foram considerados os diferentes pesos para os termos, mencionados anteriormente e descritos na Tabela 1. As Tabelas 2 e 3 mostram os resultados dos testes efetuados com as diferentes implementações para a seleção das melhores frases após a decomposição da matriz em valores singulares, destacando os melhores resultados para cada métrica ROUGE.

No caso dos métodos baseados na decomposição de matrizes não negativas, foram consideradas apenas 10 iterações para efetuar a decomposição da matriz. Foram testadas diver-

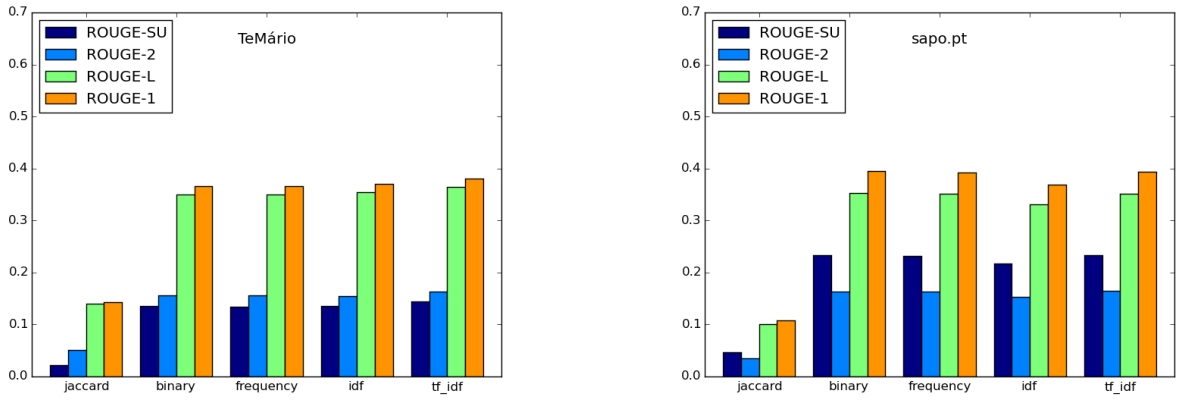


Figura 1: Resultados para os corpora TeMário (figura à esquerda) e *sapo.pt* (figura à direita), para as várias métricas ROUGE e quando usando o algoritmo TextRank com probabilidades iniciais uniformes.

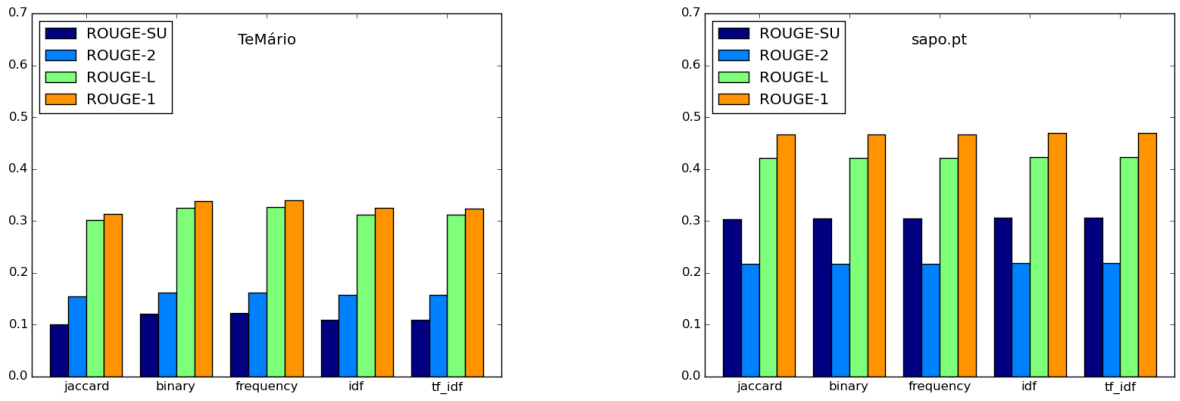


Figura 2: Resultados para os corpora TeMário (figura à esquerda) e *sapo.pt* (figura à direita), para as várias métricas ROUGE e quando usando o algoritmo TextRank com probabilidades iniciais não-uniformes.

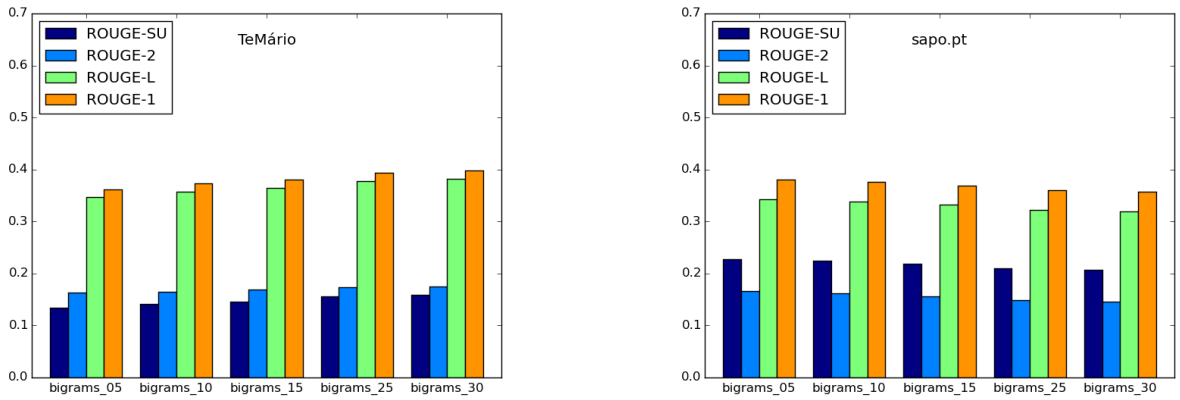


Figura 3: Resultados para os corpora TeMário (figura à esquerda) e *sapo.pt* (figura à direita), para as várias métricas ROUGE e quando usando a abordagem baseada no número de bi-gramas da lista de k mais frequentes.

Peso	Fórmula
Peso Binário (BN)	$\text{BN}(j,i) = 0\|1$
Frequência (TF)	$\text{TF}(j,i) =$ frequência do termo i no documento j
Número Inverso de Frases (IDF)	$\text{IDF}(j,i) = \log(\text{nº de frases / nº de frases com o termo } i)$
TF-IDF	$\text{TF-IDF}(j,i) = \text{TF}(j,i) \times \text{IDF}(j,i)$

Tabela 1: Pesos usados nas abordagens baseadas em decomposição de matrizes ou baseadas em grafos.

	LSA Classic				**LSA Squared**			
	IDF	Bin.	TF	TF-IDF	IDF	Bin.	TF	TF-IDF
ROUGE-1	0.3695	**0.4249**	0.4237	0.3719	0.2266	0.2821	**0.2882**	0.2325
ROUGE-2	0.1608	**0.1941**	0.1938	0.1634	0.0651	0.0918	**0.1009**	0.0716
ROUGE-L	0.3319	**0.3820**	0.3814	0.3341	0.2008	0.2487	**0.2565**	0.2072
ROUGE-SU	0.2295	**0.2739**	0.2732	0.2323	0.1033	0.1400	**0.1499**	0.1103

Tabela 2: Resultados para os diferentes esquemas de pesagem de termos, nos métodos baseados em decomposição de matrizes em valores singulares, para o corpus *sapo.pt.*

	LSA Classic				**LSA Squared**			
	IDF	Bin.	TF	TF-IDF	IDF	Bin.	TF	TF-IDF
ROUGE-1	0.3865	**0.4116**	0.4076	0.3860	0.4117	**0.4369**	0.4352	0.4093
ROUGE-2	0.1491	**0.1658**	0.1633	0.1486	0.1604	**0.1768**	0.1759	0.1581
ROUGE-L	0.3692	**0.3942**	0.3903	0.3700	0.3930	**0.4179**	0.4166	0.3909
ROUGE-SU	0.1422	**0.1629**	0.1612	0.1424	0.1595	**0.1832**	0.1810	0.1582

Tabela 3: Resultados para os diferentes esquemas de pesagem de termos, nos métodos baseados em decomposição de matrizes em valores singulares, para o corpus TeMário.

sas dimensionalidades para o número de tópicos usados na decomposição da matriz, tendo sido testadas as dimensões de 25%, 50%, 75% e 100% do tamanho da matriz inicial. Os resultados estão exibidos nas Tabelas 4 e 5. Quando os valores entre os tópicos são iguais, selecionámos como melhor implementação o algoritmo que continha o número de tópicos mais pequeno.

5 Análise dos Resultados

A Tabela 6 mostra os resultados obtidos após a execução dos diferentes tipos de algoritmos nos dois corpora testados. Os melhores resultados com todos os tipos de algoritmos estão assim em destaque na Tabela 6.

Os resultados obtidos demonstram que os métodos baseados na seleção da primeira frase de uma notícia obtêm melhores resultados na construção de títulos de forma extrativa, em termos das várias métricas ROUGE testadas, conforme referenciado na Tabela 6. Na Figura 4 apresentam-se as distribuições para os valores da similaridade de Jaccard entre as frases que se encontram nas primeiras n posições dos documentos do corpus *sapo.pt*, e a frase que constitui o título do documento respetivo. Como se pode ver, as frases nas primeiras posições são claramente mais semelhantes para com os títulos, justificando-se desta forma os bons resultados obtidos através desta *baseline* muito simples.

Para a geração de sumários de notícias, a abordagem que apresentou melhores resultados foi a implementação do algoritmo LSA Squared utilizando um peso binário para os termos, em relação a todas as métricas ROUGE.

Efetuando uma análise aos algoritmos que utilizaram SVD, na implementação LSA Classic obteve-se os melhores resultados utilizando uma representação binária de cada termo nos documentos. Quanto à implementação LSA Squared, os resultados foram diferentes para ambos os corpus testados, tendo sido obtidos melhores resultados no corpus do *sapo.pt* com a contabilização da frequência de cada palavra nos respetivos documentos. Quanto ao corpus TeMário, os melhores resultados foram obtidos utilizando uma representação binária das palavras nos respetivos documentos, para ambos os algoritmos. Comparando ambas as implementações, a implementação LSA Classic obteve melhores resultados na geração de títulos de notícias, enquanto a implementação LSA Squared obteve melhores resultados na geração de sumários com mais do que uma frase.

		NMF GR				**NMF ExtR**			
ROUGE	%Frases	IDF	Bin.	TF	TF-IDF	IDF	Bin.	TF	TF-IDF
ROUGE-1	25%	0.2230	**0.2987**	0.2965	0.2264	0.2286	**0.3060**	0.3056	0.2338
	50%	0.2091	0.2679	0.2671	0.2264	0.2259	0.2876	0.2919	0.2301
	75%	0.1934	0.2507	0.2535	0.1934	0.2238	0.2799	0.2853	0.2289
	100%	0.1805	0.2286	0.2339	0.1814	0.2226	0.2720	0.2783	0.2264
ROUGE-2	25%	0.0660	0.1037	**0.1064**	0.0701	0.0661	0.1048	**0.1111**	0.0722
	50%	0.0641	0.0912	0.0926	0.0650	0.0661	0.0960	0.1037	0.0713
	75%	0.0601	0.0842	0.0872	0.0603	0.0645	0.0920	0.1004	0.0703
	100%	0.0567	0.0765	0.0800	0.0572	0.0652	0.0877	0.0964	0.0696
ROUGE-L	25%	0.1982	**0.2645**	0.2642	0.2021	0.2026	0.2693	**0.2721**	0.2084
	50%	0.1883	0.2391	0.2396	0.1886	0.2009	0.2543	0.2596	0.2055
	75%	0.1742	0.2246	0.2275	0.1740	0.1988	0.2471	0.2543	0.2043
	100%	0.1635	0.2051	0.2105	0.1645	0.1978	0.2406	0.2481	0.2020
ROUGE-SU	25%	0.1029	0.1552	**0.1566**	0.1072	0.1047	0.1578	**0.1629**	0.1113
	50%	0.0978	0.1364	0.1373	0.0985	0.1043	0.1455	0.1527	0.1097
	75%	0.0896	0.1254	0.1288	0.0896	0.1025	0.1397	0.1482	0.1084
	100%	0.0834	0.1119	0.1171	0.0841	0.1022	0.1348	0.1433	0.1073

Tabela 4: Resultados para as várias medidas ROUGE quando considerando diferentes números de variáveis latentes, nos métodos utilizando a decomposição de matrizes não negativas e sobre o corpus *sapo.pt.*

		NMF GR				**NMF ExtR**			
ROUGE	%Frases	IDF	Bin.	TF	TF-IDF	IDF	Bin.	TF	TF-IDF
ROUGE-1	25%	0.3492	0.3534	0.3520	0.3496	0.3590	0.3636	0.3649	0.3606
	50%	0.3576	**0.3686**	0.3679	0.3566	0.4044	0.3927	0.3906	0.4024
	75%	0.3317	0.3579	0.3572	0.3372	**0.4049**	0.3825	0.3833	0.4029
	100%	0.3333	0.3519	0.3481	0.3334	0.3992	0.3772	0.3734	0.3997
ROUGE-2	25%	0.1312	0.1392	0.1373	0.1310	0.1346	0.1446	0.1453	0.1350
	50%	0.1473	**0.1493**	0.1473	0.1309	0.1573	**0.1595**	0.1566	0.1560
	75%	0.1225	0.1419	0.1419	0.1259	0.1559	0.1549	0.1550	0.1548
	100%	0.1255	0.1397	0.1367	0.1247	0.1535	0.1519	0.1492	0.1531
ROUGE-L	25%	0.3330	0.3374	0.3358	0.3338	0.3419	0.3467	0.3480	0.3438
	50%	0.3418	**0.3526**	0.3512	0.3407	0.3864	0.3757	0.3735	0.3846
	75%	0.3179	0.3423	0.3420	0.3233	**0.3871**	0.3659	0.3670	0.3852
	100%	0.3183	0.3359	0.3322	0.3188	0.3814	0.3605	0.3563	0.3816
ROUGE-SU	25%	0.1163	0.1204	0.1199	0.1166	0.1224	0.1279	0.1284	0.1236
	50%	0.1220	**0.1333**	0.1324	0.1209	**0.1546**	0.1495	0.1481	0.1538
	75%	0.1066	0.1243	0.1246	0.1103	0.1546	0.1440	0.1445	0.1531
	100%	0.1081	0.1210	0.1186	0.1079	0.1509	0.1395	0.1367	0.1509

Tabela 5: Resultados para as várias medidas ROUGE quando considerando diferentes números de variáveis latentes, nos métodos utilizando a decomposição de matrizes não negativas e sobre o corpus TeMário.

	Textos Jornalísticos PT-PT				**Textos PT-BR no Corpus TeMário**			
	R-1	R-2	R-L	R-SU	R-1	R-2	R-L	R-SU
TextRank	0.3955	0.1646	0.3528	0.2339	0.3807	0.1635	0.3645	0.1443
TextRank Init	0.4699	0.2185	0.4236	0.3058	0.3400	0.1620	0.3269	0.1222
LSA Classic	0.4249	0.1941	0.3820	0.2739	0.4116	0.1658	0.3942	0.1629
LSA Squared	0.2882	0.1941	0.2565	0.1499	**0.4369**	**0.1768**	**0.4179**	**0.1832**
NMF ExtR	0.3060	0.1111	0.2721	0.1629	0.4049	0.1595	0.3871	0.1546
NMF GR	0.2987	0.1064	0.2645	0.1566	0.3686	0.1493	0.3526	0.1333
Primeiras Frase(s)	**0.4701**	**0.2186**	**0.4238**	**0.3060**	0.3307	0.1620	0.3183	0.1107
Frase(s) Aleatória(s)	0.2893	0.1062	0.2770	0.0819	0.1828	0.0586	0.1656	0.0856
Bi-gramas	0.3802	0.1667	0.3420	0.2282	0.3980	0.1757	0.3815	0.1586

Tabela 6: Resultados obtidos pelos vários métodos e para ambos os corpora.

Quanto aos algoritmos baseados em decomposição de matrizes não negativas, nomeadamente a implementação NMF-GR, os resultados também foram distintos para cada um dos corpora. No corpus *sapo.pt*, as melhores implementações, envolvem a utilização de 25% do número máximo de variáveis latentes a selecionar, utilizando representações binárias das palavras nas frases e a contabilização da frequência de cada palavra nos respetivos documentos, para diferentes medidas ROUGE.

Quanto ao corpus TeMário, os melhores resultados foram obtidos usando uma representação binária, com 50% quanto ao número de variáveis latentes a usar na decomposição da matriz. Na implementação NMF-Extr, os resultados também foram diferentes para ambos os corpora testados, tendo-se obtido melhores resultados, para quase todas as medidas ROUGE, com a utilização da frequência das palavras nos respetivos documentos, com 25% no número de variáveis latentes, no corpus *sapo.pt*. No corpus do TeMário, os melhores resultados foram obtidos usando uma representação baseada no número inverso de frases, e com 75% no número de variáveis latentes a selecionar, embora não para todas as métricas ROUGE. Comparando ambas as implementações, a implementação NMF-Extr obteve melhores resultados em ambos os corpora.

Os testes efetuados com as duas variantes do algoritmo TextRank também demonstraram resultados diferentes para cada um dos corpora. Quanto à seleção de títulos de notícias, os melhores resultados foram obtidos utilizando a métrica IDF para todas as métricas ROUGE, e para seleção das melhores frases os melhores resultados foram obtidos utilizando a métrica TF para todas as métricas ROUGE. A implementação do algoritmo TextRank com probabilidades iniciais não-uniformes obteve piores resultados no corpus do TeMário e obteve melhores resultados no corpus do *sapo.pt*, tendo resultados aproximados com a abordagem de seleção da primeira frase.

Num teste separado, tentámos verificar qual a influência que o parâmetro α da Equação 9, o qual controla o decaimento da importância dada às frases que ocorrem nas posições iniciais do documento a sumarizar, tem sobre os resultados finais. Os dois gráficos da Figura 5 ilustram os resultados obtidos sobre as coleções TeMário e *sapo.pt*. Os melhores resultados correspondem aos valores $\alpha = 0.85$ no caso do *sapo.pt*, e $\alpha = 0.25$ no caso do TeMário, embora as variações nos resultados sejam muito pequenas.

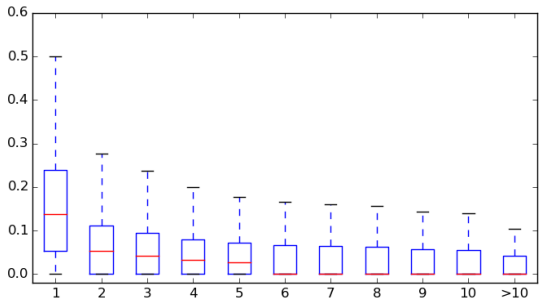


Figura 4: Distribuições para os valores da similaridade de Jaccard entre as frases do corpus *sapo.pt* que se encontram nas primeiras n posições, para com os títulos dos respetivos documentos.

Para os testes efetuados com os diferentes números de bi-gramas no método *baseline*, também se obtiveram resultados diferentes em ambos os corpora. Para a seleção dos títulos de notícias, a abordagem que obteve melhores resultados seleciona as frases que continham bigramas dos cinco bi-gramas mais frequentes, enquanto para a seleção das melhores frases para um sumário, a abordagem que obteve melhores resultados baseia-se na seleção das frases que continham o maior número de bi-gramas possíveis dentro dos trinta bi-gramas mais frequentes.

6 Conclusões e Trabalho Futuro

Este artigo apresentou uma comparação sistemática de diferentes abordagens extrativas para a tarefa de sumarizar documentos individuais correspondendo a textos jornalísticos escritos em Português. Através da utilização da bancada ROUGE como forma de medir a qualidade dos sumários produzidos, foram reportados resultados para dois domínios experimentais diferentes, envolvendo (i) a geração de títulos para textos jornalísticos escritos na variante Europeia do Português, e (ii) a geração de sumários com base em artigos jornalísticos escritos na variante Brasileira do Português. Os resultados obtidos demonstram que métodos heurísticos simples, baseados na seleção da primeira frase de uma notícia, obtêm melhores resultados na construção de títulos de forma extrativa, em termos de várias métricas ROUGE. Para a geração de sumários mais longos do que uma frase, o método que obteve melhores resultados foi o método LSA Squared, baseado na decomposição SVD de uma matriz de termos por frases.

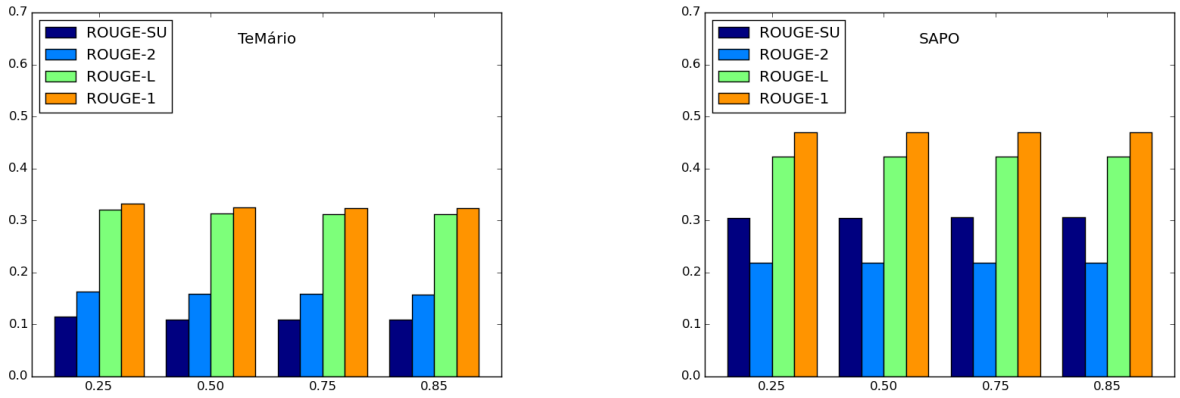


Figura 5: Resultados para os corpora TeMário (figura à esquerda) e *sapo.pt* (figura à direita), para as várias métricas ROUGE e quando usando o algoritmo TextRank com representações baseadas em TF-IDF, com probabilidades iniciais não-uniformes e com diferentes valores para o parâmetro α.

Para trabalho futuro, planeamos dar continuidade ao trabalho apresentado neste artigo, particularmente focando no problema da geração de títulos para artigos jornalísticos, ambicionando a integração de um módulo de sumarização automática, com estas características, num sistema de recomendação de notícias. Nesta aplicação em concreto, pretende-se abordar a geração de sumários curtos que não só capturem os aspetos mais importantes dos artigos, mas que também sejam personalizados em função dos interesses individuais dos utilizadores do sistema de recomendação, e que possam aumentar o rácio de cliques nas notícias apresentadas. Este é um problema muito importante no contexto de portais de notícias on-line, tais como o do serviço *sapo.pt*.

Pensamos que um método puramente extrativo terá sempre muitas limitações na aplicação concreta à geração de títulos para artigos jornalísticos e, como tal, planeamos integrar métodos de resolução de anáforas e de co-referências nas etapas de pré-processamento, por forma a poder enriquecer as frases antes de um processo de seleção para a formação de sumários. Planeamos também efetuar testes com outros métodos baseados em grafos e em adaptações do algoritmo PageRank, na tarefa de sumarização. Em particular, pensamos testar representações dos textos baseadas em grafos bi-partidos, em que os nós correspondentes a frases se interliguem com nós representando diferentes tipos de conceitos (e.g., termos individuais, entidades mencionadas, tópicos latentes, etc.) extraídos dos documentos textuais a sumarizar.

No passado, autores como Banko et al. (2000), Dorr et al. (2003) ou Alfonseca et al. (2013) abordaram já o desenvolvimento de abordagens específicas para a geração de títulos de artigos jornalísticos, indo além da sumarização extractiva. Nós pretendemos combinar abordagens para a sumarização extractiva e para a compressão de frases (e.g., consultar os trabalhos da autoria de Berg-Kirkpatrick et al. (2011) e de Almeida & Martins (2013) como exemplos recentes de abordagens deste tipo), abordando desta forma a geração de bons títulos para os artigos jornalísticos a apresentar num portal *on-line*.

Referências

Alfonseca, Enrique, Daniele Pighin & Guillermo Garrido. 2013. HEADY: News headline abstraction through event pattern clustering. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1243–1253.

Almeida, Miguel B. & André F. T Martins. 2013. Fast and robust compressive summarization with dual decomposition and multi-task learning. Em *Proceedings of the Annual Meeting of the Association for Computer Linguistics*, 196–206.

Aone, Chinatsu, Mary Ellen Okurowski, James Gorlinsky & Bjornar Larsen. 1999. A trainable summarizer with knowledge acquired from robust NLP techniques. Em Inderjeet Mani & Mark T. Maybury (eds.), *Advances in Automatic Text Summarization*, MIT Press.

Bach, Nguyen, Qin Gao, Stephan Vogel & Alex Waibel. 2011. TriS: A statistical sentence simplifier with log-linear models and margin-based discriminative training. Em *Proceedings of the International Joint Conference on Natural Language Processing*, 474–482.

Banko, Michele, Vibhu O. Mittal & Michael J. Witbrock. 2000. Headline generation based on statistical translation. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 318–325.

Barzilay, Regina, Kathleen R McKeown & Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. Em *Proceedings of the Annual Meeting of the Association for Computer Linguistics*, 550–557!

Baxendale, Phyllis B. 1958. Machine-made index for technical literature: An experiment. *IBM Journal of Research and Development* 2(4).

Berg-Kirkpatrick, Taylor, Dan Gillick & Dan Klein. 2011. Jointly learning to extract and compress. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 481–490.

Carbonell, Jaime & Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. Em *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 335–336.

Conroy, John M & Dianne P O'Leary. 2001. Text summarization via Hidden Markov Models. Em *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 406–407.

Coster, William & David Kauchak. 2011. Learning to simplify sentences using Wikipedia. Em *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, 1–9.

Dorr, Bonnie, David Zajic & Richard Schwartz. 2003. Hedge trimmer: a parse-and-trim approach to headline generation. Em *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 1–8.

Edmundson, Harold P. 1969. New methods in automatic extracting. *Journal of the ACM* 16(2).

Erkan, Günes & Dragomir R Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* 22(1).

Franceschet, Massimo. 2011. PageRank: standing on the shoulders of giants. *Communications of the ACM* 54(6).

Fung, Pascale & Grace Ngai. 2006. One story, one flow: Hidden Markov story models for multilingual multidocument summarization. *ACM Transactions on Speech and Language Processing* 3(2).

Gong, Yihong & Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. Em *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 19–25.

Hirao, Tsutomu, Jun Suzuki & Hideki Isozaki. 2009. Automatic summarization as a combinatorial optimization problem. *Transactions of the Japanese Society for Artificial Intelligence* 24(2).

Jing, Hongyan & Kathleen R McKeown. 2000. Cut and paste based text summarization. Em *Proceedings of the Conference of North American Chapter of the Association for Computational Linguistics Conference*, 178–185.

Knight, Kevin & Daniel Marcu. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence* 139(1).

Kupiec, Julian, Jan Pedersen & Francine Chen. 1995. A trainable document summarizer. Em *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 68–73.

Lee, Daniel D. & H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755).

Lee, Ju-Hong, Sun Park, Chan-Min Ahn & Daeho Kim. 2009. Automatic generic document summarization based on non-negative matrix factorization. *Information Processing & Management* 45(1).

Lin, Chin-Yew. 1999. Training a selection function for extraction. Em *Proceedings of the International Conference on Information and Knowledge Management*, 55–62.

Lin, Chin-Yew. 2004a. Looking for a few good metrics: Automatic summarization evaluation - how many samples are enough? Em *Proceedings of the NTCIR Workshop on Research in Information Access Technologies, Information Retrieval, Question Answering and Summarization*, s.p.

Lin, Chin-Yew. 2004b. ROUGE: a package for automatic evaluation of summaries. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 25–26.

Litvak, Marina, Mark Last & Menahem Friedman. 2010. A new approach to improving multilingual summarization using a genetic algorithm. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 927–936.

Luhn, Hans P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2(2).

Mani, Inderjeet, Gary Klein, David House, Lynette Hirschman, Therese Firmin & Beth Sundheim. 2002. Summac: a text summarization evaluation. *Natural Language Engineering* 8(01). 43–68.

Mashechkin, Igor, Mikhail Petrovskiy, D.S. Popov & Dmitry V. Tsarev. 2011. Automatic text summarization using latent semantic analysis. *Programming and Computer Software* 37(6).

Maziero, Erick Galani, Vinícius Rodrigues Uzêda, Tiago Salgueiro Pardo & Maria das Graças Volpe Nunes. 2007. TeMário 2006: Estendendo o córpus TeMário. Relatório Técnico. NILC-TR-07-06 Núcleo Interinstitucional de Linguística Computacional.

McKeown, Kathleen & Dragomir R Radev. 1995. Generating summaries of multiple news articles. Em *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 74–82.

Mihalcea, Rada. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. Em *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, s.p.

Mihalcea, Rada & Paul Tarau. 2005. A language independent algorithm for single and multiple document summarization. Em *Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing*, 19–24.

Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38(11).

Nenkova, Ani, Sameer Maskey & Yang Liu. 2011. Automatic summarization. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, s.p.

Osborne, Miles. 2002. Using maximum entropy for sentence extraction. Em *Proceedings of the Workshop on Automatic Summarization*, 1–8.

Page, Lawrence, Sergey Brin, Rajeev Motwani & Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Relatório Técnico. 1999-66 Stanford InfoLab.

Pardo, Thiago Alexandre Salgueiro. 2008. Sumarização automática: Principais conceitos e sistemas para o português brasileiro. Relatório Técnico. NILC-TR-08-04 Núcleo Interinstitucional de Linguística Computacional.

Pardo, Thiago Alexandre Salgueiro & Lucia Helena Machado Rino. 2003. TeMário: Um corpus para sumarização automática de textos. Relatório Técnico. NILC-TR-03-09 Núcleo Interinstitucional de Linguística Computacional.

Radev, Dragomir R, Eduard Hovy & Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Computational Linguistics* 28(4).

Radev, Dragomir R, Hongyan Jing & Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. Em *Proceedings of the Workshop on Automatic Summarization*, 21–30.

Salton, Gerard, Amit Singhal, Mandar Mitra & Chris Buckley. 1997. Automatic text structuring and summarization. *Information Processing & Management* 33(2).

Shen, Dou, Jian-Tao Sun, Hua Li, Qiang Yang & Zheng Chen. 2007. Document summarization using conditional random fields. Em *Proceedings of the International Joint Conference on Artifical Intelligence*, 2862–2867.

Siddharthan, Advaith & Kathleen McKeown. 2005. Improving multilingual summarization: using redundancy in the input to correct MT errors. Em *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 33–40.

Steinberger, Josef & Karel Ježek. 2004. Text summarization and singular value decomposition. Em *Proceedings of the International Conference on Advances in Information Systems*, 245–254.

Svore, Krysta Marie, Lucy Vanderwende & Christopher J.C. Burges. 2007. Enhancing single-document summarization by combining RankNet and third-party sources. Em *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 448–457.

Thapar, Vishal, Ahmed A Mohamed & Sanguthevar Rajasekaran. 2006. Consensus text

summarizer based on meta-search algorithms. Em *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology*, 403–407.

Wan, Xiaojun & Jianwu Yang. 2008. Multi-document summarization using cluster-based link analysis. Em *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 299–306.

Wang, Dingding & Tao Li. 2010. Many are better than one: Improving multi-document summarization via weighted consensus. Em *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 809–810.

Yamangil, Elif & Rani Nelken. 2008. Mining Wikipedia revision histories for improving sentence compression. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 137–140.

Yeh, Jen-Yuan, Hao-Ren Ke, Wei-Pang Yang & I-Heng Meng. 2005. Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing & Management* 41(1). 75–95.

Hacia una clasificación verbal automática para el español: estudio sobre la relevancia de los diferentes tipos y configuraciones de información sintáctico-semántica*

Towards an automatic verb classification for Spanish: study of the relevance of different types and configurations of syntactico-semantic information

Lara Gil-Vallejo
Universitat Oberta de Catalunya
lgilva@uoc.edu

Irene Castellón
Universitat de Barcelona
icastellon@ub.edu

Marta Coll-Florit
Universitat Oberta de Catalunya
mcollfl@uoc.edu

Jordi Turmo
Universitat Politècnica de Catalunya
turmo@lsi.upc.edu

Resumen

En este trabajo nos centramos en la adquisición de clasificaciones verbales automáticas para el español. Para ello realizamos una serie de experimentos con 20 sentidos verbales del corpus Sensem. Empleamos diferentes tipos de atributos que abarcan información lingüística diversa y un método de clustering jerárquico aglomerativo para generar varias clasificaciones. Comparamos cada una de estas clasificaciones automáticas con un gold standard creado semiautomáticamente teniendo en cuenta construcciones lingüísticas propuestas desde la lingüística teórica. Esta comparación nos permite saber qué atributos son más adecuados para crear de forma automática una clasificación coherente con la teoría sobre construcciones y cuales son las similitudes y diferencias entre la clasificación verbal automática y la que se basa en la teoría sobre construcciones lingüísticas.

Palabras clave

Clasificación verbal, clustering, construcciones

Abstract

In this work we focus on the automatic acquisition of verbal classifications for Spanish. To do so, we perform a series of experiments with 20 verbal senses that belong to the Sensem corpus. We use different kinds of features that include diverse linguistic information and an agglomerative hierarchical clustering method to generate a number of classifications. We compare each of these automatic classifications with a semi-automatically created gold standard, which is built on the basis of linguistic constructions proposed by theoretical linguistics. This comparison allows us to investigate which features are adequate to build a verb classification coherent with linguistic constructions theory and which are the similarities and differences between an automatic verbal classification and a verb classification based on the theory of linguistic constructions.

Keywords

Verb classification, clustering, constructions

1 Introducción

Los lexicones computacionales tienen un gran valor dentro del área del Procesamiento del Lenguaje Natural. VerbNet (Schuler, 2005) ha sido empleado en múltiples tareas, como por ejemplo etiquetación de papeles semánticos (Giuglea & Moschitti, 2006), elaboración de sistemas de diálogo automático (Swift, 2005) o desambiguación de sentidos verbales (Brown et al., 2014). El modelo de lexicón de VerbNet presenta además la ventaja de estar organizado por clases. Las clases verbales estructuran información relativa al verbo y a sus argumentos, lo que permite eliminar información redundante y elaborar generalizaciones (Schulte im Walde, 2006). Por ejemplo, la clase *appear-48.1.1* de VerbNet contiene 41 verbos que comparten esquemas sintáctico-semánticos, lo que permite usar los atributos asociados a la clase en tareas de Procesamiento del Lenguage Natural, generalizando la información que aporta cada verbo individualmente.

*Este trabajo ha sido realizado en el marco del proyecto Adquisición de escenarios de conocimiento a través de la lectura de textos (SKATeR, TIN2012-38584-C06-01) y gracias a una beca de investigación de la escuela de doctorado de la Universitat Oberta de Catalunya.

Sin embargo, la elaboración manual de un lexicón es costosa y requiere bastante tiempo y recursos que en ocasiones no están disponibles. Por ello, en los últimos años se han realizado varios experimentos y trabajos con el objetivo de adquirir un lexicón verbal de forma automática o semiautomática que pueda aplicarse satisfactoriamente a diversas tareas. En concreto, se ha utilizado texto anotado a diferentes niveles o lexicones de subcategorización como VALEX (Korhonen et al., 2006) para crear clasificaciones verbales automáticas asociadas a información sintáctico-semántica.

El objetivo de este trabajo es averiguar qué atributos lingüísticos son más adecuados para una clasificación sintáctico-semántica automática de verbos para el español usando técnicas de clustering, con el fin de hacer una selección de los mismos y aplicarlos posteriormente en una clasificación más amplia de unidades verbales. Para este fin, hemos realizado diversos experimentos con varias clasificaciones verbales. Para obtener estas clasificaciones verbales hemos escogido un conjunto controlado de verbos que presentan diferentes iniciadores, campos semánticos y esquemas sintácticos y hemos empleado varios tipos de atributos y un algoritmo de clustering para crear la clasificación. Los atributos contienen información lingüística sintáctico-semántica (funciones sintácticas, roles semánticos, preferencias selectivas, entre otros). Además de utilizar diferente tipo de información, hemos experimentado con diversas configuraciones de los rasgos lingüísticos y diferentes tipos de valor de los atributos. En cuanto al algoritmo, elegimos el clustering jerárquico aglomerativo, ya que es coherente con las clasificaciones verbales manuales, que son taxonómicas y que recogen la idea de la existencia de diferentes grados de similitud entre los miembros de las clases. Por otro lado, consideramos que es interesante poder observar la distribución de los sentidos verbales por clases en función del nivel de la jerarquía escogido.

2 Trabajos previos

Las clasificaciones verbales automáticas se elaboran generalmente a partir de la aplicación de un algoritmo supervisado o no supervisado a datos extraídos de un corpus. Presentan una serie de ventajas e inconvenientes sobre las manuales. Como desventaja podemos apuntar el hecho de que, al ser generadas automáticamente a partir de información de corpus, pueden contener ruido o clases no del todo coherentes, frente a la precisión que podemos encontrar en una clasificación manual. Por otro lado, las clasificaciones verbales automáticas pueden alcanzar una gran cobertura con un coste mínimo. El número de propuestas de clasificaciones verbales automáticas creció considerablemente a partir del trabajo teórico de clasificación verbal de Levin (1993), en el que se basa VerbNet, uno de los lexicones verbales más empleados en Procesamiento del Lenguaje Natural. La hipótesis de Levin es que el significado de un verbo determina su comportamiento en cuanto a la expresión e interpretación de sus argumentos. Esta hipótesis ha sido la base para muchas de las propuestas de clasificación verbal automática. Por lo tanto, la mayor parte del trabajo realizado en el área de clasificación verbal automática tiene por objetivo crear unas clases verbales similares a las que propone Levin. Para ello exploran diferentes características lingüísticas y algoritmos de clustering. A continuación ofrecemos un panorama general del trabajo realizado en esta área, tanto el que está basado en las clasificaciones verbales de Levin, como aquellas propuestas que tienen como objetivo adquirir otro tipo de clasificación verbal.

Con relación a aquellos trabajos cuyo objetivo es adquirir una clasificación similar a la de Levin (y que, por tanto, usan adaptaciones o traducciones de la clasificación de Levin como gold standard) podemos diferenciar entre aquellos que usan un enfoque supervisado y los que usan un enfoque no supervisado (clustering). Ambos tipos modelan los verbos basándose en un conjunto de características lingüísticas orientadas a capturar las alternancias de diátesis en las que Levin basa su clasificación. Sin embargo, en el caso de los enfoques no supervisados, la clase a la que pertenece un verbo no es conocida a priori.

En cuanto a los enfoques no supervisados, que será nuestra perspectiva, generalmente emplean patrones de subcategorización en combinación con diferentes algoritmos, como por ejemplo Joanis et al. (2008) y Li & Brew (2008). Los patrones de subcategorización enriquecidos con preferencias selectivas han demostrado dar lugar a una mayor precisión a la hora de inducir las clases de Levin como vemos en Sun & Korhonen (2009) y Vlachos et al. (2009)

Este método para realizar clasificaciones verbales también se ha empleado para otras lenguas diferentes del inglés. Para evaluar estas clasificaciones se han empleado diferentes métodos: Brew & Schulte im Walde (2002) y Schulte im Walde (2006) crean un gold standard manual para el alemán, mientras que Falk et al. (2012) construyen automáticamente una base de datos para el francés con criterios similares a la de VerbNet.

Otra alternativa común es la de traducir las clases de Levin, lo que permite una comparación entre los resultados en ambos idiomas. Sun & Korhonen (2009) obtienen para el francés una medida-F de 54.6 (la medida-F para el equivalente inglés es de 80.4). Scarton et al. (2014) obtienen una medida-F de 42.77 para el portugués brasileño. En ambos casos los atributos que mejor funcionan son los patrones de subcategorización enriquecidos con preferencias selectivas y preposicionales. Para el español, Ferrer (2004) aplica un clustering jerárquico a 514 verbos y los evalúa con la clasificación manual de Vázquez et al. (2000). Usa probabilidades de diferentes tipos de patrones de subcategorización, obteniendo una medida Rand de 0.07 para 15 clusters.

Una aproximación diferente es el trabajo de Sun et al. (2013), que no emplea patrones de subcategorización, sino que propone un método alternativo para capturar las alternancias de diátesis de los verbos, basándose en la idea de que una alternancia de diátesis puede aproximarse calculando la probabilidad conjunta de dos patrones de subcategorización.

Entre aquellos trabajos que se apartan del objetivo de adquirir una clasificación verbal similar a la de Levin, podemos mencionar la propuesta de Merlo & Stevenson (2001), que utiliza un enfoque supervisado para clasificar verbos en tres grupos: inacusativos, inergativos y de objeto nulo. Finalmente, cabe mencionar también el trabajo de Lenci (2014), cuyo objetivo es descubrir clases verbales. Para ello usa patrones de subcategorización y preferencias selectivas en un corpus del italiano, empleando uno de estos patrones de subcategorización como semilla para después hacer particiones según rasgos más específicos entre los verbos que lo contienen.

En general los trabajos mencionados asignan los lemas verbales a una sola clase, lo que no permite dar cuenta de la polisemia verbal. Este factor puede ser muy importante, ya que la mayoría de los verbos tienen al menos dos sentidos. Al modelar un verbo sin tener en cuenta sus sentidos puede obtenerse un modelo poco preciso, ya que en realidad la mayor parte de la información se obtiene del sentido más frecuente, mientras que aquellos sentidos menos frecuentes quedan sin modelar o distorsionan el modelo (Korhonen et al., 2003).

3 Metodología

A continuación explicamos la metodología que hemos seguido en este trabajo. En primer lugar, detallamos los criterios para seleccionar los sentidos verbales para los experimentos (3.1). Seguidamente explicamos el proceso de creación de un gold standard (3.2), tomando construcciones lingüísticas propuestas desde la lingüística teórica. El gold standard es una referencia con la que se pueden comparar las clasificaciones verbales automáticas para comprobar si se obtienen clases equivalentes. A continuación (3.3), explicamos el proceso de extracción de información lingüística del corpus para generar los datos que sirven de base para los experimentos. Además, en este mismo apartado explicamos el tipo de algoritmo de clustering que empleamos para elaborar las diferentes clasificaciones verbales automáticas.

3.1 Selección de sentidos verbales

En nuestro trabajo hemos optado por realizar experimentos con sentidos verbales, en vez de lemas, para obtener modelos más precisos. En concreto, trabajamos con un único sentido por verbo, esto es, no incluimos pares polisémicos para poder modelizar sin ambigüedad. No obstante, reconocemos que el fenómeno de la polisemia verbal es algo que se ha de tener en cuenta y tratar en cualquier aplicación computacional.

Se escogen 20 sentidos verbales del corpus Sensem (Fernández-Montraveta & Vázquez, 2014) que aparecen con una frecuencia mayor de 10 frases en el corpus para asegurar la representatividad de las diferentes propiedades sintáctico-semánticas asociadas con los sentidos. Estos 20 sentidos presentan diferentes esquemas sintácticos, pertenecen a diferentes campos semánticos, correspondientes a los supersenses de Wordnet asociados a los synsets del Multilingual Central Repository (Gonzalez-Agirre & Rigau, 2013) y poseen diferentes tipos de iniciadores del evento: causativos, agentivos y experimentadores. Estas tres características permiten que el conjunto escogido sea representativo, pese al limitado número de sentidos verbales. A continuación mostramos la clasificación de los sentidos seleccionados según el campo semántico al que pertenecen:[1]

- estado: parecer_1, valer_1, estar_14.
- comunicación: valorar_2, explicar_1.
- cognición: gustar_1, pensar_2.
- movimiento: perseguir_1, viajar_1,volver_1, montar_2.
- cambio: abrir_18, cerrar_19, crecer_1, morir_1.

[1]Para una definición de los sentidos verbales y número de ocurrencias de cada uno en el corpus se puede consultar el anexo B

- percepción: ver_1, escuchar_1.
- actividad (social y corporal): trabajar_1, dormir_1, gestionar_1.

3.2 Creación del gold standard

A continuación detallamos el proceso de creación del gold standard, que es una clasificación verbal basada en propuestas teóricas sobre construcciones. Definimos la noción de *construcción* como un signo lingüístico, con forma y significado, que comprende estructura sintáctica y roles semánticos. Esta definición es coincidente con la noción de construcción de Goldberg (1994) y la de diátesis de Levin. Esta clasificación servirá para evaluar las clasificaciones creadas automáticamente, lo que permitirá escoger los atributos adecuados para crear una clasificación automática similar a una clasificación manual, mucho más costosa de realizar.

El primer paso en la creación del gold standard es seleccionar los atributos lingüísticos que configurarán las clases verbales. En nuestro caso hemos utilizado estructuras sintácticas básicas descritas en múltiples gramáticas como Barreto & Bosque (1999). Además, hemos empleado construcciones adaptadas de Levin, teniendo en cuenta los trabajos de Cifuentes Honrubia (2006) y Vázquez et al. (2000). Tomamos las construcciones aisladas, es decir, cada uno de los pares en una alternancia de diátesis, lo que no impone restricciones sobre el tipo de alternancia en el que participan los verbos. Dado que la cantidad de sentidos escogidos es limitada para controlar el efecto de los diferentes atributos, se han preferido aquellas construcciones que tienen un carácter más general sobre aquellas específicas para determinados verbos. A continuación listamos y explicamos brevemente estas estructuras y construcciones. Empleamos como atributos cinco estructuras sintácticas básicas: transitiva, intransitiva, ditransitiva, predicativa y atributiva; además, contamos con trece construcciones:

1. Causativa prototípica: Construcción en la que se explicita la causa de un evento por medio de un sujeto. El sujeto puede ser un agente (volitivo) o una causa (no volitiva). El objeto está afectado por el evento en diversos grados. Ej. *La falta de lluvias secó el río*
2. Anticausativa prototípica (con "se"): Es una construcción intransitiva donde la entidad afectada ocupa la posición de sujeto. Ej. *El río se secó*
3. Causativa de perífrasis: Es una causativa en la que el predicado aparece en infinitivo junto con el auxiliar "hacer". Ej. *Los fuertes vientos han hecho bajar las temperaturas*
4. Anticausativa sin "se": el constituyente que expresa la causa se elide. Una entidad no afectada ocupa la posición de sujeto. Ej. *Las temperaturas han bajado*
5. Voz media: Expresa un estado o propiedad del sujeto sin combinarse con un verbo atributivo. Generalmente van con un complemento adverbial que refuerza la lectura estativa, a diferencia de la anticausativa prototípica, que tiene una interpretación dinámica. Ej. *La pintura se esparce con facilidad.*
6. Impersonal pronominal: El verbo aparece en tercera persona, no tienen sujeto gramatical explícito ni recuperable por el contexto. Ej. *Se aconseja el uso obligatorio del cinturón*
7. Sujeto oblicuo: El iniciador del evento aparece en una posición encabezada por una preposición. Se suele subdividir en varios tipos, pero dado que nuestro número de ejemplos es pequeño, no hemos tenido en cuenta estas subdivisiones. Ejs. *La gente se beneficia de las nuevas medidas*
8. Reflexiva: La acción expresada por el sujeto recae sobre sí mismo. Ej. *María se peina.*
9. Recíproca: El sujeto de estas construcciones es plural. Cada uno de los componentes del sujeto ejerce una acción sobre los otros, a la vez que la recibe de los demás. Ej. *Juan y Pedro se desafiaron.*
10. Pasiva perifrástica: El objeto ocupa una posición topicalizada y el verbo se construye con un auxiliar. Generalmente el agente se puede expresar mediante un sintagma preposicional. Ej. *Los bizcochos fueron comidos por los niños*
11. Pasiva refleja: Se construye con la partícula "se". El sujeto se pospone a la partícula. El iniciador de la acción no se explicita pero suele ser agentivo. Ej. *Se pasaron los trabajos a ordenador*
12. Objeto cognado: El objeto que mantiene una relación etimológica con el verbo, por ello las frases con esta construcción tienen un sentido tautológico. Ej. *Cantamos una canción*
13. Resultativa con "estar": Detalla el estado resultado de la acción expresada por el verbo. Ej. *El pan está cortado*

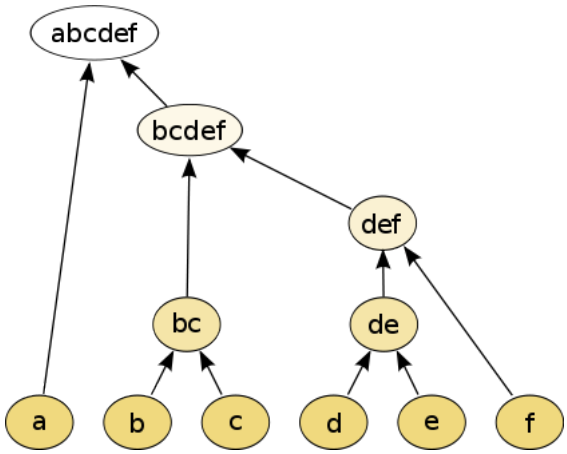


Figura 1: Modelización del clustering jerárquico aglomerativo (adaptado de la entrada de agrupamiento jerárquico de Wikipedia)

Para crear las clases verbales del gold standard aplicamos un clustering jerárquico aglomerativo junto con estos atributos y los sentidos verbales descritos. En el clustering jerárquico aglomerativo cada elemento (en nuestro caso sentidos verbales) pertenece inicialmente a un grupo. En cada paso se van fusionando los dos grupos con menor distancia (ver figura 1). La distancia entre dos grupos se calcula aplicando una función de distancia entre algunos de sus elementos (por ejemplo, distancia euclidea, distancia del coseno, etc.). La selección de dichos elementos se puede realizar de formas diferentes, que se definen como tipos de enlace. En nuestros experimentos hemos empleado cuatro tipos de enlace distintos (simple, completo, promedio y promedio ponderado) para comprobar el efecto que tienen en las agrupaciones de los verbos, con lo cual obtenemos un gold standard para cada tipo de enlace.

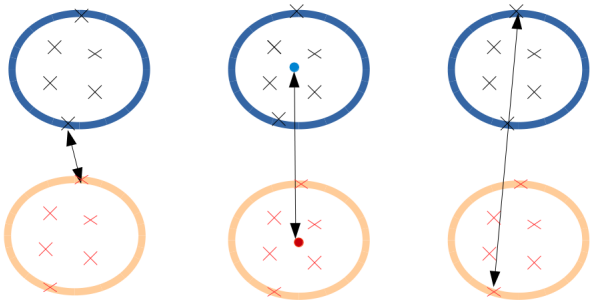


Figura 2: Enlace simple, promedio y completo

En la figura 2 podemos ver cómo se calcula la distancia entre grupos en los tres tipos de enlace: de izquierda a derecha mostramos el enlace simple, el enlace promedio (el promedio ponderado es una variante de este) y el enlace completo. En el enlace simple la distancia entre dos grupos viene dada por la mínima distancia entre los elementos de ambos grupo. En el enlace promedio la distancia entre dos grupos se calcula como promedio de la distancia entre todos los pares de elementos de ambos grupos. En el enlace promedio ponderado la distancia entre dos grupos se define de la misma manera que en el caso del enlace promedio, pero se tienen en cuenta los grupos previos que pasaron a formar parte de los grupos actuales. Finalmente, en el enlace completo la distancia entre dos grupos se calcula teniendo en cuenta los elementos más dispares de ambos grupos.

El resultado del clustering jerárquico es una jerarquía de posibles agrupaciones, cada una de ellas definida por cada nivel de la jerarquía. Una vez obtenido el resultado del clustering, se debe decidir el nivel de agrupación más apropiado. Para ello, tres lingüistas evaluaron las distintas agrupaciones que contenían de 4 a 10 clases y finalmente, después de varias reuniones de discusión, se llegó al acuerdo de que el modelo de 6 clases era el más adecuado, ya que en él hay una serie de agrupaciones de los sentidos verbales comunes para los cuatro tipos de enlace que son coherentes con la teoría lingüística. Aparecen siempre en la misma clase los sentidos de carácter estativo *estar_14* y *parecer_1*. En otra clase aparecen juntos *abrir_18*, *cerrar_19*, *crecer_1* y *morir_1*, que son verbos que expresan cambio (junto con ellos aparece también *dormir_1*, que se trata de una actividad). También juntos en una clase se agrupan *escuchar_1*, *explicar_1*, *gestionar_1*, *perseguir_1*, *ver_1* y *valorar_2*, que generalmente tienen iniciadores humanos u organizaciones. *Trabajar_1* y *volver_1*, intransitivos agentivos, también permanecen juntos en todos los tipos de enlace y en ocasiones se agrupan con otros verbos. *Valer_1* y *gustar_1* siempre son miembros únicos de su grupo. Los demás alternan entre los grupos ya mencionados. Las clases resultantes pueden consultarse en la columna izquierda del anexo A.

3.3 Experimentación

En los experimentos se emplean diferentes atributos lingüísticos extraídos del corpus Sensem:

- atributos semánticos de los argumentos:
  - Roles semánticos obtenidos a partir de un mapping jerárquico realizado entre los roles de Sensem y la propuesta de Lirics (Bonial et al., 2011):
    - roles semánticos finos (40 roles),
    - roles semánticos abstractos (16 roles);

  - Supersenses de Wordnet (Miller, 1995) (45 supersenses);
  - Ontología de SUMO (Niles & Pease, 2003) (1000 términos). Los supersenses y los términos de la ontología de SUMO se obtienen a partir del núcleo de los argumentos verbales, que en Sensem están anotados con synsets.
- atributos morfosintácticos: función sintáctica; categoría morfológica; construcción, que recoge aspectos como la topicalización o destopicalización del sujeto lógico, la reflexividad o la impersonalidad.
- aspecto oracional (estado, evento, proceso).

Mediante la selección de esta información configuramos diferentes espacios de atributos para los experimentos. Con el fin de obtener una representación lo más completa posible de los predicados, cada atributo semántico se combina con uno sintáctico. Por otro lado, para explorar el rol del aspecto, que no se ha tenido en cuenta generalmente a la hora de elaborar clasificaciones verbales automáticas, realizamos una versión de estos atributos combinada con el aspecto de las frases. Finalmente, para valorar el potencial de los roles semánticos a la hora de definir una clasificación verbal, añadimos otro atributo que consiste en roles semánticos sin combinarlos con información sintáctica.

Como resultado tenemos 27 tipos de atributos según el tipo de información lingüística que recogen (por ejemplo, *sintaxis+supersenses*, *sintaxis+roles de sensem*, *categoría morfosintáctica+ontología SUMO+aspecto*, etc). A su vez, estos atributos admiten tres configuraciones diferentes de información: rasgos aislados, constituyentes y patrones. En la figura 3 presentamos un ejemplo de anotación de la frase en Semsem *Remedios abrió su bolso.*

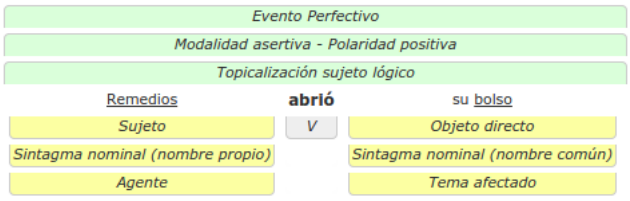


Figura 3: Anotación de una frase en el corpus Sensem

Para esta frase, con información lingüística relativa a roles y funciones sintácticas, obtendríamos las siguientes configuraciones:

- rasgos aislados (4 atributos): sujeto, agente, objeto directo, tema afectado
- constituyentes (2 atributos): sujeto-agente, objeto directo-tema afectado
- patrones (1 atributo): sujeto-agente+objeto directo-tema afectado

Por lo tanto, los 20 sentidos tomados de Sensem quedan caracterizados por los atributos sintáctico-semánticos de las frases en las que participan. En cuanto al valor del atributo, experimentamos con dos tipos diferentes: binarios (0/1) y probabilidades. Los atributos binarios toman valores 0 o 1 dependiendo de si para un sentido y un atributo dados (por ejemplo *abrir_18* y *sujeto-agente+objeto-tema*) hay al menos una frase que recoja ambos (1) o no la hay (0). Por otro lado, las probabilidades expresan, para un sentido y un atributo dados (por ejemplo *abrir_18* y *sujeto-agente+objeto-tema*), la proporción de frases en las que aparecen ambos en relación a las frases en las que participa el sentido verbal.

Para cada set de datos con un tipo de atributo obtenido mediante las combinaciones que acabamos de explicar, realizamos clustering jerárquico aglomerativo con cada uno de los cuatro enlaces posibles. Las funciones de distancia empleadas son dos: una basada en el coeficiente Dice (Dice, 1945), ya que es adecuada y ampliamente empleada para atributos binarios y otra basada en el Coseno para los probabilísticos, también muy utilizada en estos casos. Estas medidas se emplean para calcular la distancia entre dos elementos en función de los valores asociados a los atributos que los caracterizan. El número de clases deseadas en el resultado del clustering que compararemos con el gold standard se establece en un rango entre 4 y 10.

4 Evaluación y resultados

Para evaluar los resultados, comparamos cada gold standard correspondiente a un tipo de enlace con los resultados de los experimentos para este enlace. En las tablas 1, 2, 3 y 4 mostramos los resultados para cada tipo de enlace. Para cada variación de atributo-valor (rasgos aislados, constituyentes y patrones con valores probabilísticos y binarios) se muestra el número de clases y el tipo de información lingüística que conforman la clasificación automática más similar al respectivo gold standard. La similitud entre el gold standard y el resultado de cada experimento se mide empleando la información mutua ajustada, que da cuenta de la similitud entre dos etiquetados diferentes para los mismos datos. En nuestro caso, los dos etiquetados son las clases verbales definidas en el gold standard y las clases obtenidas automáticamente con datos de Sensem. La me-

dida de información mutua ajustada oscila entre 0 si las clases verbales son independientes y 1 si las clases verbales son idénticas. Hay una gran variedad de medidas de evaluación externa para algoritmos de clustering. Hemos elegido esta medida porque no presenta sesgos en cuanto al número de clases, al contrario que otras medidas muy utilizadas como la de pureza (Manning et al., 2008) y está ajustada, es decir, en el caso de una agrupación aleatoria de sentidos verbales, el valor de la medida de información mutua ajustada es 0 (Strehl, 2002).

5 Análisis de resultados

Si observamos globalmente los resultados correspondientes a todos los tipos de enlace, vemos que la información lingüística que en más ocasiones contribuye a generar una clasificación similar a la del gold standard es la combinación de supersenses y funciones sintácticas, con una información mutua ajustada media de 0.530 (este tipo de información obtiene mejores resultados en cuatro ocasiones para el enlace completo, dos para el simple y una para el de tipo promedio ponderado). Sin embargo, la combinación de roles abstractos más funciones sintácticas, que es la segunda información lingüística que más frecuentemente aparece en las tablas (dos veces para el enlace promedio, una para el completo, una para el simple y dos para el promedio ponderado) tiene una información mutua ajustada media ligeramente mayor: 0.542. En general observamos que las funciones sintácticas aparecen en muchos de los atributos que mejores resultados obtienen.

Si nos centramos en el tipo de valor, vemos que las probabilidades dan lugar a una información mutua ajustada media mayor que los atributos binarios: 0.55 frente a 0.49. En cuanto a la configuración de los atributos, las configuraciones que generalmente dan lugar a una clasificación más similar a la del gold standard son las de patrones y constituyentes, ambas con una información mutua ajustada media de 0.54. Los rasgos aislados obtienen peores resultados, con un 0.49.

En conjunto, la configuración que mejores resultados arroja es la que contiene información acerca de los supersenses y la función sintáctica organizada en patrones y con valores probabilísticos. Este tipo de atributos y valores en el enlace simple obtiene una medida de información mutua de 0.647. También cabe destacar que la combinación *roles abstractos+aspecto+función sintáctica* obtiene una de las mejores medidas, 0.627, lo que pone de relieve la importancia del aspecto como información relevante a la hora de crear una clasificación verbal automática.

Como hemos visto en el apartado de trabajos previos, hay una clasificación verbal automática para el español realizada por Ferrer (2004), que consigue una medida Rand ajustada de 0.07 clasificando 514 verbos en 15 grupos. Para tener una referencia, calculamos la medida Rand ajustada de la clasificación verbal generada por la configuración que obtiene una mayor información mutua ajustada. La medida Rand de esta clasificación es de 0.619. Pese a que se trata de un valor notablemente más alto que el que alcanza Ferrer (2004), hay que tener en cuenta que el tipo de gold standard es diferente y la cantidad de verbos es menor en nuestro caso, lo que limita el posible ruido que se generaría con un número mayor de sentidos. Aunque ambas clasificaciones no son directamente comparables, consideramos que los resultados que hemos obtenido son prometedores y nos animan a seguir trabajando en esta línea.

En lo relativo a las clases que se obtienen haciendo clustering con los datos de Sensem, vemos como en las cuatro mejores agrupaciones, una por enlace[2], hay unos rasgos comunes: de forma similar a lo que ocurre en el gold standard, estar y parecer se mantienen en una misma clase que tampoco contiene ningún otro miembro. Por el contrario, el grupo de verbos que expresaban cambio junto con *dormir_1* no se mantiene. En concreto, *abrir_18* y *cerrar_19* generalmente aparecen en un grupo separado de *crecer_1* y *dormir_1*. En este punto coinciden con la distinción hecha por Levin & Hovav (1995) entre verbos de cambio de estado que expresan un evento de causa externa y aquellos que expresan un evento de causa interna. En Levin & Hovav (1995) se definen los eventos de causa interna como aquellos en los que el argumento que acompaña al verbo posee una propiedad que es responsable del evento denotado (por ejemplo 'la planta creció') y los eventos de causa externa como aquellos en los que hay una causa externa que tiene el control del evento (por ejemplo 'la puerta se abrió'), que además puede ser hecha explícita en una construcción transitiva (por ejemplo 'el viento abrió la puerta'). *Explicar_1*, *escuchar_1*, *gestionar_1*, *perseguir_1*, *valorar_2* y *ver_1*, que aparecían siempre en el mismo grupo en el gold standard, independientemente del tipo de enlace, se mantienen juntos también en todos los enlaces de las clases obtenidas con datos de corpus. *Valer_1* aparece como único miembro de su grupo en todos los casos, tanto en el gold standard como en los grupos creados a partir de corpus.

[2]En negrita en las tablas 1-4, los verbos que componen estas clases están en la columna izquierda de las tablas del anexo A.

Configuración de los atributos	**Valor de los atributos**	**Información lingüística de los atributos**	**Número de grupos**	**Información Mutua Ajustada**
rasgos aislados	binario	SUMO aspecto funciones sintácticas	7	0.425
rasgos aislados	probabilidades	roles abstractos construcciones	6	0.598
constituyentes	binario	roles abstractos funciones sintácticas	6	0.591
constituyentes	probabilidades	roles abstractos aspecto funciones sintácticas	6	**0.627**
patrones	binario	roles abstractos morfología	6	0.598
patrones	probabilidades	roles abstractos funciones sintácticas	7	0.609

Tabla 1: Enlace promedio.

Configuración de los atributos	**Valor de los atributos**	**Información lingüística de los atributos**	**Número de grupos**	**Información Mutua Ajustada**
rasgos aislados	binario	SUMO aspecto morfología	7	0.389
rasgos aislados	probabilidades	supersenses funciones sintácticas	8	0.488
constituyentes	binario	supersenses funciones sintácticas	5	0.519
constituyentes	probabilidades	supersenses funciones sintácticas	7	0.479
patrones	binario	roles abstractos funciones sintácticas	6	0.422
patrones	probabilidades	supersenses funciones sintácticas	8	**0.551**

Tabla 2: Enlace completo.

Respecto a las diferencias entre el gold standard y las clases obtenidas, vemos que *viajar_1* y *trabajar_1* aparecen siempre juntos, mientras que en el gold standard *trabajar_1* aparecía siempre junto con *volver_1*. *Gustar_1*, que en las clases del gold standard aparecían como único miembro de su grupo, aparece en una ocasión en el mismo grupo que *crecer_1*. El resto de los verbos alternan entre dos grupos principales en las clasificaciones hechas con datos de Sensem: *pensar_2* alterna entre el grupo de *escuchar_1* y aislado, *montar_2* alterna entre *volver_1* y aislado, *volver_1* alterna entre el grupo de *montar_2* y aislado. Finalmente, *morir_1* alterna entre el grupo de *abrir_18* y el de *crecer_1*, lo que no es consecuente con el criterio de causa externa e interna, ya que de mantenerse este criterio en la clasificación automática debería permanecer con *crecer_1*.

6 Conclusiones

En este trabajo hemos analizado parámetros relevantes a la hora de hacer clasificaciones verbales automáticas empleando clustering jerárquico aglomerativo. Para ello hemos creado un gold standard para cada tipo de enlace de forma semiautomática, utilizando atributos motivados en la teoría lingüística. Posteriormente hemos realizado varios experimentos empleando diferentes tipos de parámetros y hemos analizado los resultados.

En concreto, para el clustering jerárquico aglomerativo, comprobamos que los diferentes tipos de enlace tienen un efecto en la configuración de las clases. En cuanto al diseño de los atributos, hemos visto como la configuración en patrones y

Configuración de los atributos	**Valor de los atributos**	**Información lingüística de los atributos**	**Número de grupos**	**Información Mutua Ajustada**
rasgos aislados	binario	SUMO aspecto funciones sintácticas	6	0.567
rasgos aislados	probabilidades	roles abstractos aspecto funciones sintácticas	5	0.590
constituyentes	binario	roles abstractos funciones sintácticas	6	0.561
constituyentes	probabilidades	supersenses funciones sintácticas	6	0.561
patrones	binario	SUMO aspecto funciones sintácticas	6	0.561
patrones	probabilidades	supersenses funciones sintácticas	6	**0.647**

Tabla 3: Enlace simple.

Configuración de los atributos	**Valor de los atributos**	**Información lingüística de los atributos**	**Número de grupos**	**Información Mutua Ajustada**
rasgos aislados	binario	SUMO aspecto funciones sintácticas	6	0.372
rasgos aislados	probabilidades	roles abstractos	7	0.479
constituyentes	binario	supersenses funciones sintácticas	9	0.468
constituyentes	probabilidades	roles abstractos funciones sintácticas	6	0.532
patrones	binario	roles abstractos morfología	5	0.503
patrones	probabilidades	roles abstractos funciones sintácticas	7	**0.539**

Tabla 4: Enlace promedio ponderado.

constituyentes ofrece unos resultados mejores que los rasgos aislados. Si tenemos en cuenta la mejor clasificación por enlace, son los patrones los que mejor funcionan, algo que va en la línea de los trabajos previos. En relación con esto, hemos observado que el tipo de valor que recoge de forma más efectiva la información proporcionada por los datos son las probabilidades de co-ocurrencia de verbo y atributo.

En cuanto a la información lingüística, hemos comprobado que las funciones sintácticas tienen un papel fundamental, y que ofrecen buenos resultados combinadas con roles semánticos abstractos o los supersenses de Wordnet. Además hemos demostrado que el aspecto, que generalmente no se ha tenido en cuenta en los trabajos previos, es un rasgo útil. Una inspección manual de las clases nos ha permitido observar la existencia de similitudes básicas globales entre el gold standard y las clases elaboradas con datos de Sensem.

En definitiva, en este trabajo hemos evaluado qué tipo de información sintáctico-semántica es más relevante para una clasificación automática verbal del español, así como el tipo de valor y configuración de los atributos más adecuados, empleando un conjunto acotado y controlado de sentidos verbales. Esto nos ha permitido hacer un estudio de los cambios en la configuración de las clases según el empleo de diferentes parámetros. A partir de estos resultados, el próximo paso será aplicar los parámetros obtenidos a una clasificación más amplia de unidades verbales del español.

A Clases verbales: gold standard y clase más similar

Gold standard	Clasificación más similar
1: estar_14 parecer_1 **2**: escuchar_1 explicar_1 gestionar_1 pensar_2 perseguir_1 valorar_2 ver_1 viajar_1 **3**: abrir_18 cerrar_19 crecer_1 dormir_1 montar_2 morir_1 **4**: trabajar_1 volver_1 **5**: gustar_1 **6**: valer_1	**1**: estar_14 parecer_1 **2**: valer_1 **3**: montar_2 volver_1 **4**: abrir_18 cerrar_19 morir_1 **5**: crecer_1 dormir_1 gustar_1 **6**: escuchar_1 explicar_1 gestionar_1 pensar_2 perseguir_1 trabajar_1 valorar_2 ver_1 viajar_1

Tabla 5: Enlace promedio.

Gold standard	Clasificación más similar
1: estar_14 parecer_1 **2**: trabajar_1 volver_1 **3**: gustar_1 **4**: escuchar_1 explicar_1 gestionar_1 perseguir_1 valorar_2 ver_1 **5**: abrir_18 cerrar_19 crecer_1 dormir_1 montar_2 morir_1 pensar_2 viajar_1 **6**: valer_1	**1**:abrir_18 cerrar_19 **2**: estar_14 parecer_1 **3**: crecer_1 dormir_1 morir_1 trabajar_1 viajar_1 volver_1 **4**: escuchar_1 explicar_1 gestionar_1 perseguir_1 valorar_2 ver_1 **5**: valer_1 **6**: pensar_2 **7**: montar_2 **8**: gustar_1

Tabla 6: Enlace completo.

Gold standard	Clasificación más similar
1: estar_14 parecer_1 **2**: trabajar_1 volver_1 **3**: abrir_18 cerrar_19 crecer_1 dormir_1 escuchar_1 explicar_1 gestionar_1 morir_1 pensar_2 perseguir_1 valorar_2 ver_1 viajar_1 **4**: montar_2 **5**: gustar_1 **6**: valer_1	**1**: estar_14 parecer_1 **2**: abrir_18 cerrar_19 crecer_1 dormir_1 escuchar_1 explicar_1 gestionar_1 morir_1 perseguir_1 trabajar_1 valorar_2 ver_1 viajar_1 volver_1 **3**: montar_2 **4**: valer_1 **5** pensar_2 **6**: gustar_1

Tabla 7: Enlace simple.

Gold standard	Clasificación más similar
1: estar_14 parecer_1 **2**: escuchar_1 explicar_1 gestionar_1 perseguir_1 valorar_2 ver_1 **3**: abrir_18 cerrar_19 crecer_1 dormir_1 montar_2 morir_1 pensar_2 viajar_1 **4**: trabajar_1 volver_1 **5**: gustar_1 **6**: valer_1	**1**: estar_14 parecer_1 **2**: valer_1 **3**: crecer_1 dormir_1 morir_1 trabajar_1 viajar_1 volver_1 **4‘**: abrir_18 cerrar_19 escuchar_1 explicar_1 gestionar_1 perseguir_1 valorar_2 ver_1 **5**: montar_2 **6**: gustar_1 **7**: pensar_2

Tabla 8: Enlace promedio ponderado.

B Definición de los sentidos verbales

Entre paréntesis se indica el número de ocurrencias en el corpus.
abrir_18: Descorrer el pestillo o cerrojo, desechar la llave, levantar la aldaba o desencajar cualquier otra pieza o instrumento semejante con que se cierra algo. (15)
cerrar_19: Asegurar con cerradura, pasador, pestillo, tranca u otro instrumento, una puerta, ventana, tapa, etc., para impedir que se abra. (14)
crecer_1: Incrementar la cantidad o la importancia de algo, desarrollarse. (116)
dormir_1: Permanecer en un estado en el cual todos los movimientos voluntarios son suspendidos, generalmente para descansar. (18)
escuchar_1: Poner atención a lo que se oye. (107)
estar_14: Encontrarse alguien o algo en un estado determinado. (101)
explicar_1: Aclarar algo, dar información sobre un asunto. (106)
gestionar_1: Realizar un trámite para la consecución de una cuestión. (36)
gustar_1: Encontrar atractivo o agradable alguna cosa o a alguien. (117)
montar_2: Subirse alguien en un animal o un vehículo. (26)
morir_1: Fallecer, dejar de existir algo o alguien. (115)
parecer_1: Aparentar algo, sin serlo necesariamente. (51)
pensar_2: Usar la mente alguien para examinar una idea, razonar. (25)
perseguir_1: Ir detrás de alguien o algo para alcanzarle. (53)
trabajar_1: Emplearse en cualquier ejercicio, obra, trabajo o ministerio. (80)
valorar_2: Admitir la importancia de un hecho, cosa o acción. (70)
valer_1: Tener algo un determinado valor. (45)
ver_1: Recibir una imagen a través de la vista. (86)
viajar_1: Ir de un lugar a otro que suele estar distante, generalmente mediante algún medio de transporte. (111)
volver_1: Dirigirse hacia el lugar donde ya se ha estado. (84)

Referencias

Barreto, Violeta Demonte & Ignacio Bosque. 1999. *Gramática descriptiva de la lengua española*. Espasa Calpe.

Bonial, Claire, William Corvey, Martha Palmer, Volha V Petukhova & Harry Bunt. 2011. A hierarchical unification of lirics and verbnet semantic roles. En *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, 483–489. IEEE.

Brew, Chris & Sabine Schulte im Walde. 2002. Spectral clustering for german verbs. En *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 117–124. Association for Computational Linguistics.

Brown, Susan Windisch, Dmitriy Dligach & Martha Palmer. 2014. Verbnet class assignment as a wsd task. En *Computing Meaning*, 203–216. Springer.

Cifuentes Honrubia, JL. 2006. Alternancias verbales en español. *Revista Portuguesa de Humanidades* 10. 107–132.

Dice, Lee R. 1945. Measures of the amount of ecologic association between species. *Ecology* 26(3). 297–302.

Falk, Ingrid, Claire Gardent & Jean-Charles Lamirel. 2012. Classifying french verbs using french and english lexical resources. En *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, 854–863. Association for Computational Linguistics.

Fernández-Montraveta, Ana & Gloria Vázquez. 2014. The sensem corpus: an annotated corpus for spanish and catalan with information about aspectuality, modality, polarity and factuality. *Corpus Linguistics and Linguistic Theory* 10(2). 273–288.

Ferrer, Eva Esteve. 2004. Towards a semantic classification of spanish verbs based on subcategorisation information. En *Proceedings of the ACL 2004 workshop on Student research*, 13. Association for Computational Linguistics.

Giuglea, Ana-Maria & Alessandro Moschitti. 2006. Semantic role labeling via framenet, verbnet and propbank. En *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 929–936. Association for Computational Linguistics.

Goldberg, A. 1994. *Constructions, a construction grammar approach to argument structure*. Chicago, Il: Chicago University Press.

Gonzalez-Agirre, Aitor & German Rigau. 2013. Construcción de una base de conocimiento léxico multilíngüe de amplia cobertura: Multilingual central repository. *Linguamática* 5(1). 13–28.

Joanis, Eric, Suzanne Stevenson & David James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering* 14(03). 337–367.

Korhonen, Anna, Yuval Krymolowski & Ted Briscoe. 2006. A large subcategorization lexicon for natural language processing applications. En *Proceedings of LREC*, vol. 6, .

Korhonen, Anna, Yuval Krymolowski & Zvika Marx. 2003. Clustering polysemic subcategorization frame distributions semantically. En *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, 64–71. Association for Computational Linguistics.

Lenci, Alessandro. 2014. Carving verb classes from corpora. *Word Classes: Nature, typology and representations* 332. 17.

Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Levin, Beth & Malka Rappaport Hovav. 1995. *Unaccusativity: At the syntax-lexical semantics interface*, vol. 26. MIT Press.

Li, Jianguo & Chris Brew. 2008. Which are the best features for automatic verb classification. En *ACL*, 434–442.

Manning, Christopher D, Prabhakar Raghavan, Hinrich Schütze et al. 2008. *Introduction to information retrieval*, vol. 1. Cambridge University Press Cambridge.

Merlo, Paola & Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics* 27(3). 373–408.

Miller, George A. 1995. WordNet: a lexical database for English. *Communications of the ACM* 38(11). 39–41.

Niles, Ian & Adam Pease. 2003. Mapping wordnet to the sumo ontology. En *Proceedings of the ieee international knowledge engineering conference*, 23–26.

Scarton, Carolina, Lin Sun, Karin Kipper-Schuler, Magali Sanches Duran, Martha Palmer & Anna Korhonen. 2014. Verb clustering for brazilian portuguese. En *Computational Linguistics and Intelligent Text Processing*, 25–39. Springer.

Schuler, Karin Kipper. 2005. *Verbnet: A broad-coverage, comprehensive verb lexicon*: University of Pennsylvania. Tese de Doutoramento.

Strehl, Alexander. 2002. Relationship-based clustering and cluster ensembles for high-dimensional data mining.

Sun, Lin & Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. En *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, 638–647. Association for Computational Linguistics.

Sun, Lin, Diana McCarthy & Anna Korhonen. 2013. Diathesis alternation approximation for verb clustering. En *ACL (2)*, 736–741.

Swift, Mary. 2005. Towards automatic verb acquisition from verbnet for spoken dialog processing. En *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, 115–120.

Vázquez, Gloria, Ana Fernández & M. Antònia Martí. 2000. Clasificación verbal. *Alternancias de diátesis. Quaderns de Sintagma* 3.

Vlachos, Andreas, Anna Korhonen & Zoubin Ghahramani. 2009. Unsupervised and constrained dirichlet process mixture models for verb clustering. En *Proceedings of the workshop on geometrical models of natural language semantics*, 74–82. Association for Computational Linguistics.

Schulte im Walde, Sabine. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics* 32(2). 159–194.

Estudio de la influencia de incorporar conocimiento léxico-semántico a la técnica de Análisis de Componentes Principales para la generación de resúmenes multilingües

Studying the influence of adding lexical-semantic knowledge to Principal Component Analysis technique for multilingual summarization

Óscar Alcón
Universidad de Alicante
oalcon@dlsi.ua.es

Elena Lloret
Universidad de Alicante
elloret@dlsi.ua.es

Resumen

El objetivo de la generación automática de resúmenes es reducir la dimensión de un texto y a su vez mantener la información relevante del mismo. En este artículo se analiza y aplica la técnica de Análisis de Componentes Principales, que es independiente del idioma, para la generación de resúmenes extractivos mono-documento y multilingües. Dicha técnica se estudiará con el objetivo de poder evaluar su funcionamiento cuando se incorpora (o no) conocimiento léxico-semántico, a partir del uso de recursos y herramientas dependientes del idioma. La experimentación propuesta se ha realizado en base a dos corpus de diferente naturaleza: noticias periodísticas y artículos de la Wikipedia en tres idiomas (alemán, español e inglés) para verificar el uso de esta técnica en varios escenarios. Los enfoques propuestos presentan resultados muy competitivos comparados con generadores de resúmenes multilingües existentes, lo que indica que, aunque exista un claro margen de mejora respecto a la técnica y el tipo de conocimiento incorporado, ésta tiene una gran potencial para ser aplicada en otros contextos e idiomas.

Palabras clave

PCA, Análisis de Componentes Principales, generación de resúmenes, multilingües, extractivos, entidades nombradas, identificación de conceptos

Abstract

The objective of automatic text summarization is to reduce the dimension of a text keeping the relevant information. In this paper we analyse and apply the language-independent Principal Component Analysis technique for generating extractive single-document multilingual summaries. This technique will be studied to evaluate its performance with and without adding lexical-semantic knowledge through language-dependent resources and tools. Experiments were conducted using two different corpora: newswire and Wikipedia articles in three languages (English, German and Spanish) to validate the use of this technique in several scenarios. The proposed approaches show very competitive results compared to multilingual available systems, indicating that, although there is still room for improvement with respect to the technique and the type of knowledge to be taken into consideration, this has great potential for being applied in other contexts and for other languages.

Keywords

PCA, Principal Component Analysis, automatic summarization, multilingüal summarization, extractive summarization, NER, concept identification

1 Introducción

Actualmente, el tratamiento y gestión de la información es una tarea difícil de abordar para el ser humano. En un contexto donde cada vez la cantidad de información y la heterogeneidad de la misma aumentan a un ritmo considerable, cobran una mayor importancia las técnicas automáticas de análisis y reducción de volumen para la detección y extracción de la información relevante.

Además, se dispone de una gran cantidad de información por lo que se hace necesario el desarrollo de técnicas de Procesamiento de Lenguaje Natural (PLN) para poder procesar, clasificar, extraer y resumir la información del texto. Respecto a la tarea de generación de resúmenes, cuyo objetivo es obtener una versión reducida del documento o documentos fuentes, reduciendo su contenido pero sin perder información clave (Spärck Jones, 2007), no siempre resulta sencillo determinar qué información es la más rele-

vante, debido a la variedad de factores que podemos tener en cuenta (por ejemplo, preferencias del usuario, necesidades de información, finalidad del resumen, etc.). Así, se han establecido diferentes tipologías de resúmenes (Mani & Maybury, 1999; Spärck Jones, 2007). Entre los tipos más comunes se encuentra la distinción entre resúmenes *mono-documento* (el resumen se genera a partir de un único documento de entrada) vs. *multi-documento* (varios documentos de entrada); *extractivos* (el resumen simplemente se limita a realizar una selección de las frases más relevantes) vs. *abstractivos* (el resumen contiene información expresada de distinta manera con respecto al documento fuente); *genéricos* vs. *centrados en un tema concreto*; así como también *monolingües* vs. *multilingües*, si el resumidor funciona únicamente para un idioma o para varios.

Para abordar la generación automática de los distintos tipos de resúmenes anteriormente comentados se han utilizado distintas técnicas (Nenkova & McKeown, 2011): desde técnicas superficiales que determinan la relevancia de información según el peso de las unidades que forman las frases, como por ejemplo la frecuencia de palabras y aproximaciones derivadas (McCargar, 2005), hasta enfoques que se basan en el uso de técnicas de análisis del discurso, que implican un procesamiento más profundo del texto. En este último caso encontramos como ejemplo, la técnica de las cadenas léxicas (Barzilay & Elhadad, 1999), o la técnica de la estructura retórica del discurso (RST) (Uzêda et al., 2010).

En el contexto actual, donde no solamente hay grandes cantidades de información y el ritmo de crecimiento de la misma es exponencial, sino que dicha información está disponible en una gran variedad de idiomas, es necesario investigar en técnicas de generación de resúmenes multilingües que consigan determinar la información clave sea cuál sea el idioma en el que se haya escrito. Para ello, es necesario o bien recurrir a técnicas totalmente independientes del idioma como la frecuencia de términos (Teng et al., 2008), o bien que la técnica o el conocimiento aplicado estén disponibles para varios idiomas. Una técnica independiente del idioma es el Análisis de Componentes Principales (Principal Component Analysis, PCA), que se puede aplicar a la detección y extracción de palabras clave en un texto. Dada la naturaleza de la misma, esta técnica puede ser adecuada para la generación de resúmenes multilingües, y por tanto, será la que estudiaremos en este trabajo.

Por consiguiente, el principal objetivo de este articulo es analizar la técnica PCA para la generación de resúmenes extractivos monodocumento y multilingües. Esta técnica se analizará y evaluará por un lado de forma independiente (técnica base), y por otro con el enriquecimiento mediante la incorporación de conocimiento léxico-semántico, obtenido a partir del reconocimiento de entidades nombradas (Named Entity Recognition, NER) y la identificación de conceptos sinónimos, con el fin de medir la influencia de estas técnicas sobre la técnica base, y determinar si pueden ser beneficiosas en el proceso de generación de resúmenes multilingües diseñado. Además, una vez obtenidas las palabras clave a partir del uso de la técnica PCA en sus diversas variantes, se proponen y analizan cuatro heurísticas para la selección de las frases relevantes del documento fuente, dando lugar a distintos tipos de resúmenes extractivos.

El artículo se estructura del siguiente modo: la sección 2 recoge los trabajos realizados hasta el momento acerca de resúmenes multilingües y del uso de la técnica PCA para resumir textos. En la sección 3 se explica el método implementado para la generación de resúmenes mono-documento y multilingües con PCA. La sección 4 alberga la información de los corpus empleados para la experimentación y de las medidas de evaluación que se utilizarán en la sección 5. En la sección 6 se recogen y analizan los resultados que serán comparados con sistemas previos en la sección 7. Finalmente, se exponen las conclusiones obtenidas en la sección 8.

2 Estado de la cuestión

La generación de resúmenes multilingües es una tarea aún en desarrollo dada la dificultad de poder abarcar las características particulares de cada idioma.

En (Gupta & Lehal, 2010) se recoge una serie de enfoques en relación a la tarea de resúmenes multilingües entre los que se incluye (Cowie et al., 1998), quienes parece que iniciaron la investigación en esta temática cuando presentaron MINDS, un sistema que incluye soporte para la creación de resúmenes de documentos en inglés, español, ruso y japonés. El núcleo del sistema se generó empleando técnicas como análisis estadístico, sintáctico y de estructura de documentos. Más tarde, Hovy y Lin (1997) se adentraron en la materia con SUMMARIST, un sistema que permite la generación de resúmenes tanto abstractivos como extractivos en distintos idiomas, empleando técnicas de procesamiento del lenguaje natural, junto con bases de conocimiento. En (Patel et al., 2007) se propuso también

un método, independiente del idioma, para resumir textos de distintos idiomas. Estaba basado en factores estadísticos y de estructura del documento, tales como posición en el texto o identificación de nombres comunes y propios. No obstante, utilizaba el lexema de las palabras y filtraba el documento para eliminar las palabras carecientes de contenido léxico-semántico (stopwords). El sistema se testó con documentos en inglés, hindi, gujarati y urdu. En (Lloret & Palomar, 2011) se analizaron tres enfoques distintos, empleando: 1) técnicas independientes del idioma; 2) técnicas específicas de cada idioma; y 3) aplicando traducción automática a resúmenes monolingües. Dichos enfoques se orientaron a la producción de resúmenes extractivos en cuatro idiomas diferentes (español, inglés, alemán y frances).

La competición bienal MultiLing[1] se creó en 2011 con motivo de fomentar el trabajo sobre la generación de resúmenes multilingües. Se presentan enfoques de distintos equipos de investigación para mostrar el estado del arte en la materia y enfocar las investigaciones futuras (Giannakopoulos et al., 2011; Kubina et al., 2013). En uno de los trabajos de esta competición se presentó el sistema MUSE (Litvak & Last, 2013). Para el desarrollo de este sistema emplearon un algoritmo genético para la optimización lineal de diversas medidas de clasificación de frases. Otro enfoque fue presentado en (Conroy et al., 2013), donde se describía el uso del Análisis de la Semántica Latente (Latent Semantic Analysys, LSA) para la generación de resúmenes multi-documento para 10 idiomas distintos. Cabe destacar como en la última edición de MultiLing (Multiling 2013), algunos de los sistemas participantes alcanzaron unos resultados similares a los obtenidos con resúmenes manuales (Giannakopoulos, 2013). Concretamente, el sistema WBU (Steinberger, 2013) fue el que mejor resultados consiguió, basándose en la técnica LSA. Este sistema fue probado en 10 idiomas, y en los que en 5 de ellos quedó en primera posición.

La técnica PCA, al igual que la técnica LSA, se encuentra englobada dentro de las técnicas de minería de datos, pero se diferencian en la manera de calcular la matriz, teniendo menos dispersión en el caso de la técnica PCA. Esta técnica se ha utilizado con anterioridad para la tarea de resumir texto en Lee, Kim y Park (2003), donde se propuso un sistema para la extracción de frases de un texto que representen la información relevante, y se empleó la técnica PCA para extraer las palabras clave del documento, seleccionando las frases según la cantidad de palabras clave que incluían, siendo la más relevante aquella que albergara mayor cantidad de palabras clave. Obtuvieron buenos resultados (medida F = 0.416) para textos en coreano. Vikas et al., (2008) desarrollaron otro enfoque en donde se expone un sistema para resumir textos mono-documento y multi-documento, utilizando un Modelo Espacial de Vectores Semánticos (Semantic Vector Space Model, SVSM) para modelar el conjunto de documentos. La técnica PCA se empleó para extraer características referentes al tema del documento sobre un conjunto de textos de distinta temática en inglés. Más recientemente, la técnica PCA se ha aplicado con éxito a la generación automática de *hashtags*, en la que se logran unos resultados cercanos al 60 % (Estellés Arolas et al., 2010). Los *hashtags* de un tuit se pueden equiparar a las palabras clave que pueden resumir el texto expresado en un tuit, y por tanto, se demuestra la utilidad de la técnica PCA para los nuevos géneros textuales surgidos con la Web 2.0.

Revisados los trabajos previos en éste ámbito, la técnica PCA no ha sido investigada con anterioridad para generar resúmenes multilingües, a pesar de ser una técnica independiente del idioma, ni tampoco se ha analizado la influencia de incorporar información léxico-semántica al proceso. En base a esto, nuestra contribución en este artículo es el estudio de dicha técnica para la producción de resúmenes extractivos monodocumento y multilingües, así como el análisis de la influencia de incorporar conocimiento léxico-semántico, dependiente del idioma, a la técnica base. En nuestra propuesta, la técnica PCA se utilizará para extraer las palabras clave de los textos, que serán posteriormente empleadas para escoger las frases mas relevantes que conformarán el cuerpo del resumen final.

3 Aplicación de la técnica PCA para la generación de resúmenes multilingües

Los procesos de generación de resúmenes automáticos se han caracterizado por seguir un flujo genérico que engloba tres fases claramente diferenciadas (Sparck-Jones, 1999): i) interpretación; ii) transformación; y iii) generación de resúmenes.

Partiendo de esa base se ha formulado una propuesta de método para la generación de resúmenes aplicando la técnica PCA para textos multilingües, reflejado en la Figura 1. Como se puede observar, la fase de interpretación (sección 3.1) será en la cual se realice un preprocesado para obtener la información de interés y prepararla para la siguiente fase. En la fase de trans-

[1] http://multiling.iit.demokritos.gr/

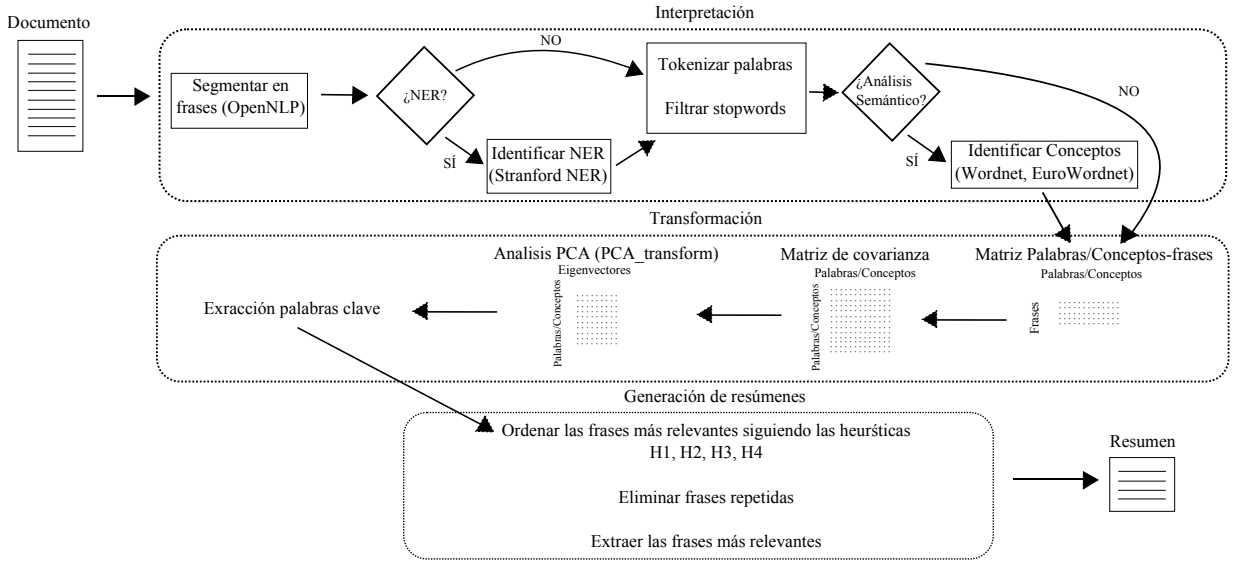


Figura 1: Flujo de acción de nuestro método de generación de resúmenes basado en la técnica PCA.

formación (sección 3.2) será en donde se aplique la técnica PCA para obtener las palabras clave del documento mediante el procesamiento de la información del texto. Finalmente, en la fase de generación de resúmenes (sección 3.3) se definen una serie de heurísticas para la selección y extracción de las frases más relevantes y formar el resumen final.

3.1 Interpretación

El método desarrollado tiene como entrada un texto al cual se le aplica el siguiente preprocesamiento lingüístico: i) segmentación en frases (OpenNLP[2]); ii) tokenización; iii) eliminación de palabras carecientes de contenido semántico (stopwords). Se incluye la opción de añadir conocimiento léxico-semántico mediante la identificación de entidades nombradas, explicado en la sección 3.1.1, y/o la identificación de conceptos, posteriormente explicado en la sección 3.1.2.

3.1.1 Reconocimiento de Entidades Nombradas

El Reconocimento de Entidades Nombradas (Named Entity Recognition, NER) consiste en etiquetar el texto de entrada para reconocer secuencias de palabras que sean nombres de personas, organizaciones y lugares, llamadas entidades (Tjong et al., 2003).

Para nuestro enfoque utilizamos como reconocedor de entidades la herramienta *Stanford Named Entity Recognizer*[3] que funciona para varios idiomas (español, inglés y alemán, entre otros).

[2] https://opennlp.apache.org/

[3] http://nlp.stanford.edu/software/CRF-NER.shtml

3.1.2 Identificación de conceptos

La identificación de conceptos en nuestro enfoque se basa en la detección de sinónimos, considerando cada conjunto de sinónimos como un único concepto y agrupando sus apariciones a lo largo del texto. Para ello, se ha utilizado *WordNet* (Miller, 1995) y *EuroWordnet* (Vossen, 2004) ya que estos recursos recogen conocimiento léxico-semántico para distintos idiomas - *Wordnet* para inglés, y *EuroWordnet* para un conjunto de idiomas europeos (entre ellos, el español y el alemán) - agrupando las palabras por conjuntos de sinónimos y almacenando las relaciones entre los mismos. La razón por la que se utilizó EuroWordnet frente a recursos que pudieran estar más actualizados para cada idioma, como Mutlilingual Central Repository[4] para el castellano o el GermaNet[5] fue para validar en una primera versión de la investigación realizada si el uso de este tipo de conocimiento integrado en la técnica PCA era apropiado o no.

Para identificar los conceptos, nos basamos en el sentido más frecuente como algoritmo de desambiguación[6], puesto que la relación resultados-coste computacional es aceptable en el estado de la cuestión (los resultados de esta aproximación están alrededor del 50 % (McCarthy, 2011)). Por tanto, en esta fase, se utiliza esta aproximación para buscar el primer synset de cada palabra en el documento. *Wordnet* y *EuroWordnet* estiman como primer synset de cada

[4] http://adimen.si.ehu.es/web/MCR

[5] http://www.sfs.uni-tuebingen.de/GermaNet/index.shtml

[6] La tarea de desambiguación del sentido de las palabras no es objeto de este artículo.

palabra el significado más frecuente de la misma y por tanto, el más probable. Si dos palabras tienen el mismo primer synset, serán consideradas como sinónimos y sus apariciones en el texto serán sumadas. Por ejemplo, los términos *"devastar"* y *"destrozar"*, aunque distintos en su morfología, se considerarían englobados en el mismo concepto, puesto que el primer synset de ambas palabras es el *"00260311"*.

3.2 Transformación: la técnica PCA para la detección de palabras clave

la técnica PCA se basa en un algoritmo matemático que reduce la dimensionalidad de los datos conservando la mayor parte de la variación en el conjunto de datos (Ringnér, 2008). Ante un gran volumen de datos con distintas variables, el objetivo de este algoritmo es encontrar una serie de patrones o tendencias dentro del conjunto de datos de entrada y transformar linealmente el conjunto de variables original en un conjunto considerablemente menor de variables incorreladas (Dunteman, 1989).

3.2.1 Generación de la matriz de componentes principales

A partir de la información obtenida en la etapa anterior, en primer lugar se genera la matriz de palabras/conceptos-frases, en la que para cada palabra/concepto se recoge el número de ocurrencias de esa palabra/concepto en el texto. A partir de esta matriz, se genera la matriz de covarianza para descifrar las relaciones existentes entre las palabras/conceptos del texto. Esta matriz será utilizada para obtener las componentes principales (eigenvectores) y su correspondiente valor propio (eigenvalue) mediante la aplicación de la técnica PCA, utilizando la librería de Java PCA_transform[7].

La aplicación de la técnica PCA devuelve una matriz en la cual las columnas son los eigenvectores (ordenados en orden descendiente determinado por el eigenvalue asociado) y las filas serían las variables (que este caso serían las palabras/conceptos del texto). Cada eigenvector se conforma con la contribución de cada variable, que determina la importancia de dicha variable en el eigenvector. De cada eigenvector extraemos la palabra(s) o concepto(s) que presente mayor contribución, considerándolas como palabras clave del texto, que posteriormente serán empleadas con el fin de seleccionar las frases más relevantes, explicado en la sección 3.3.

[7] https://github.com/mkobos/pca_transform

3.3 Generación de resúmenes

En esta fase se define la estrategia para escoger las frases en función de los valores obtenidos de la técnica PCA. Con las palabras clave extraídas, se plantean diferentes heurísticas para la extracción de las frases más relevantes del texto para la realización del resumen automático:

H1: Se selecciona, por orden de aparición en el texto, la frase que contiene el término extraído del eigenvector, realizándose este proceso para todas las palabras clave determinadas por la técnica PCA. En el caso de que la frase ya esté seleccionada, se escogería la siguiente frase donde aparece.

H2: Se selecciona, por orden de aparición en el texto, sólo la primera frase que contiene el término extraído del eigenvector, prosiguiendo para todas las palabras clave. En el caso de que la frase ya esté seleccionada se pasaría al siguiente término. Aunque esta heurística es similar a la anterior, la principal diferencia entre ambas radica en el número de frases que se pueden incluir para cada término extraído y el tratamiento de las frases que ya han sido incluidas anteriormente para formar parte del resumen. Mientras que en H1 se puede seleccionar más de una frase que contenga el término, si la frase que se va a seleccionar ya se ha incluido en el resumen anteriormente, según la estrategia definida para H2, sólo seleccionamos la primera frase que contiene el término y cuando nos encontramos con una frase que ya ha sido incluida, se finaliza con ese término y se pasa al siguiente para seguir con el proceso de selección de frases.

H3: Se seleccionan todas las frases donde aparecen las palabras clave extraídas, siguiendo el orden determinado por la técnica PCA.

H4: Se buscan las frases en las que aparecen las palabras clave y se escogen aquellas frases en las que haya incluidos al menos dos términos. Serán ordenadas en el resumen según la importancia de dichas palabras clave incluidas.

Las frases seleccionadas para cada heurística serán las conformantes del resumen final, eliminando previamente las frases repetidas. Cabe mencionar que si existen dos o más palabras con el mismo valor máximo dentro de un mismo eigenvector, se extraería las frases correspondientes para cada palabra. Del mismo modo y en el caso de utilizarse conocimiento léxico-semántico, cuando un concepto está representado por varios

sinónimos, se extraería las frases correspondientes para cada sinónimo.

4 Entorno de evaluación

En esta sección se explican los corpus que serán utilizados para testar el funcionamiento del sistema (sección 4.1). Además, los resúmenes generados serán evaluados mediante las medidas determinadas en la sección 4.2.

4.1 Corpus

Para probar el funcionamiento de la técnica desarrollada se han empleado dos corpus multilingües de distinta naturaleza: corpus JRC y corpus de entrenamiento de MultiLing 2015 y concretamente, nos vamos a centrar en 3 idiomas: inglés, español y alemán.

4.1.1 Corpus JRC

El corpus JRC[8] dispone de un conjunto de noticias periodísticas en 7 idiomas, donde para cada idioma hay 20 documentos agrupados en 4 temas: genética; conflicto Israel-Palestina; malaria; ciencia y sociedad. Así mismo, proporciona un conjunto de resúmenes humanos que se emplearán como modelos para la evaluación, contando con 4 resúmenes modelo extractivos para cada uno de los documentos. Los documentos en inglés, español y alemán tienen de media 820, 927 y 836 palabras, respectivamente.

4.1.2 Corpus de entrenamiento de MultiLing 2015

El corpus de entrenamiento de MultiLing 2015[9] está formado por artículos extraídos de la Wikipedia, donde para cada idioma hay 30 documentos. Se cuenta también con un resumen modelo abstractivo para cada documento, que puede ser empleado para realizar una evaluación automática. Los documentos en inglés, español y alemán tienen de media 3973, 6311 y 4248 palabras, respectivamente.

4.2 Medidas de evaluación

La evaluación de los resúmenes se realiza de forma cuantitativa, centrándonos exclusivamente en el contenido de los resumenes generados, y para ellos utilizamos la herramienta ROUGE (Lin, 2004), por ser una de las más utilizadas en este campo. Esta herramienta permite la evaluación automática de resúmenes mediante la comparación del número de n-gramas coincidentes de dichos resúmenes con respecto a unos resúmenes modelo. Partiendo de esta premisa, ROUGE implementa diferentes métricas, teniendo en cuenta: la similitud de unigramas (ROUGE-1); la similitud de bigramas (ROUGE-2); la secuencia común más larga (ROUGE-L) y la similitud de bigramas evitando unigramas (ROUGE-SU4). Además, para cada uno de los indicadores antes mencionados, ROUGE devuelve las siguientes medidas: Precisión, Recall y medida F.

5 Experimentación

En esta sección se describen los experimentos realizados sobre el método propuesto para comprobar y analizar su funcionamiento. Nuestro objetivo es determinar la idoneidad de la técnica PCA aplicada a resúmenes multilingües y analizar la influencia de la incorporación de conocimiento léxico-semántico de forma gradual. La experimentación se va a realizar para tres idiomas (inglés, español y alemán), puesto que existen recursos que nos permiten obtener el tipo de información léxico-semántica que necesitamos.

Para cada una de las cuatro heurísticas formuladas, se plantean los siguientes enfoques para la incorporación de conocimiento léxico-semántico de forma gradual:

(a) *PCA_base*: no se utiliza conocimiento léxico-semántico y se incluyen en la matriz obtenida a partir de la técnica PCA todas las palabras del documento (excepto stopwords).

(b) *PCA_base+CPT*: se incorpora al método base la identificación de conceptos, de tal manera que se incluyen en la matriz obtenida a partir de la técnica PCA todas las palabras y los conceptos (excepto stopwords).

(c) *PCA_base+CPT+NER*: se enriquece el método anterior con un proceso de NER, y por tanto, se incluyen en la matriz obtenida a partir de la técnica PCA todas las palabras, los conceptos y las NER identificadas (excepto stopwords).

6 Resultados y discusión

Los resultados de los experimentos realizados se muestran en la Tabla 1 y la Tabla 2, correspon-

[8] http://optima.jrc.it/Resources/2010_JRC_multilingual-summary-evaluation.zip.

[9] http://users.iit.demokritos.gr/~ggianna/MultiLing2015/multilingMss2015Training.tar.gz

		(a) PCA_base				(b) PCA_base+CPT				(c) PCA_base+CPT+NER			
		R-1	R-2	R-L	R-SU4	R-1	R-2	R-L	R-SU4	R-1	R-2	R-L	R-SU4
Inglés	H1	0.54200	0.33613	0.51832	0.36249	**0.54576**	**0.33841**	**0.52220**	**0.36528**	0.56159	0.36577	0.53736	0.39008
	H2	0.53714	0.32581	0.51184	0.35312	0.54101	0.32523	0.51509	0.35496	0.54760	0.34377	0.52402	0.36977
	H3	**0.57797**	0.39244	**0.55502**	0.41626	0.53348	0.33485	0.50800	0.36249	0.56370	0.37574	0.53888	0.40023
	H4	0.57774	**0.39444**	0.55475	**0.41821**	0.52967	0.33088	0.50299	0.35851	**0.56614**	**0.37775**	**0.54167**	**0.40269**
Español	H1	0.58221	0.37045	0.55315	0.39808	**0.59845**	**0.39288**	**0.57178**	**0.41868**	0.59827	**0.39197**	**0.57054**	0.41400
	H2	0.58105	0.36419	0.55157	0.39277	0.59449	0.38367	0.56499	0.41041	**0.59878**	0.39173	0.57024	**0.41580**
	H3	**0.58786**	**0.38836**	**0.56577**	**0.41942**	0.58710	0.38716	0.56487	0.41842	0.57595	0.37670	0.55273	0.40714
	H4	0.58850	0.38738	0.56571	0.41780	0.58357	0.38151	0.56046	0.41247	0.57913	0.37658	0.55423	0.40712
Alemán	H1	0.52992	0.35991	0.50932	0.36927	0.52527	0.35010	0.50142	0.36028	0.52156	0.35081	0.49985	0.35968
	H2	**0.53300**	**0.36391**	**0.51279**	**0.37096**	**0.53364**	**0.36400**	**0.50996**	**0.37128**	**0.54070**	**0.37632**	**0.51825**	**0.38161**
	H3	0.49642	0.32025	0.47508	0.33613	0.50753	0.33865	0.48716	0.35300	0.51446	0.34086	0.49254	0.35533
	H4	0.49835	0.32219	0.47705	0.33787	0.50750	0.33725	0.48654	0.35138	0.52221	0.35199	0.50156	0.36509

Tabla 1: Resultados ROUGE corpus JRC (Medida F)

		(a) PCA_base				(b) PCA_base+CPT				(c) PCA_base+CPT+NER			
		R-1	R-2	R-L	R-SU4	R-1	R-2	R-L	R-SU4	R-1	R-2	R-L	R-SU4
Inglés	H1	**0.43991**	**0.11488**	**0.34974**	**0.16889**	0.43633	**0.11173**	0.34399	0.16504	0.43255	0.10784	0.34147	0.16328
	H2	0.43812	0.11062	0.34873	0.16589	**0.43785**	0.11116	**0.34496**	**0.16552**	**0.43633**	**0.11044**	**0.34607**	**0.16603**
	H3	0.41745	0.09944	0.33273	0.15724	0.41749	0.10070	0.33370	0.15799	0.40548	0.09234	0.32157	0.14997
	H4	0.41866	0.09941	0.33315	0.15751	0.41757	0.10068	0.33316	0.15803	0.40510	0.09219	0.32094	0.14981
Español	H1	0.43477	**0.12031**	0.35097	0.17376	0.43940	**0.11739**	0.35430	**0.17348**	0.43491	0.11601	0.34709	0.17016
	H2	**0.44031**	0.11863	**0.35482**	**0.17461**	**0.44024**	0.11617	**0.35614**	0.17259	**0.44125**	**0.12012**	**0.35578**	**0.17509**
	H3	0.41901	0.10849	0.33738	0.16820	0.41978	0.10963	0.33900	0.16925	0.41949	0.10567	0.33620	0.16499
	H4	0.41878	0.10833	0.33692	0.16809	0.42030	0.10995	0.33937	0.16951	0.42004	0.10601	0.33600	0.16539
Alemán	H1	**0.26939**	**0.04422**	**0.19030**	**0.07323**	**0.27202**	**0.04266**	**0.19197**	**0.07352**	**0.26873**	0.04078	**0.19231**	**0.07103**
	H2	0.26782	0.04403	0.19020	0.07242	0.26662	0.04111	0.18846	0.07135	0.25946	**0.04116**	0.18835	0.06889
	H3	0.25689	0.03099	0.18424	0.06621	0.25689	0.03099	0.18424	0.06621	0.25641	0.03017	0.18391	0.06576
	H4	0.25780	0.03099	0.18462	0.06638	0.25684	0.03100	0.18414	0.06622	0.25636	0.02996	0.18380	0.06553

Tabla 2: Resultados ROUGE corpus de entrenamiento de MultiLing 2015 (Medida F)

dientes al corpus JRC y corpus de entrenamiento de MultiLing 2015 respectivamente. Resulta lógico que los resultados del corpus JRC sean notablemente superiores a los del corpus MultiLing 2105, debido a la naturaleza distinta de cada corpus y a los resúmenes utilizados como modelo para la evaluación, dado que los del corpus JRC son de tipo extractivo, mientras que los del MultiLing 2015 son abstractivos.

La aportación de conocimiento léxico-semántico es muy dependiente de los propios textos, de la cantidad de entidades y de los sinónimos que alberguen. Es por ello que se han estudiado los dos corpus para ver las características de los documentos (número de palabras por documento (sin stopwords); número de NER identificadas; y número de conceptos sinónimos identificados).

Estas características, reflejadas en la Tabla 3, nos pueden servir para sacar conclusiones de la relevancia de la adición de conocimiento léxico-semántico al proceso. Cabe destacar la escasa identificación de conceptos sinónimos en los corpus, siendo los documentos de entrenamiento del Multiling en español en los que más sinónimos se han identificado (1.76 %). A partir de un análisis más en profundidad de los corpus utilizados, se ha comprobado que efectivamente no abunda el uso de conceptos sinónimos y por el contrario, predominan más las referencias a entidades nombradas, sobre todo a entidades de tipo lugar (Inglaterra, Egipto, Japón, Estados Unidos, Europa, entre otras) y de tipo persona (Jane Austen, Harris Bigg-Wither, Thomas Blanchard, Woo-Suk Hwang, Presidente Bush, entre otras).

	Idioma	**PPD**	**NER**	**Conceptos sinónimos**
JRC	Inglés	372.10	3.52 %	0.48 %
	Español	454.15	3.33 %	0.40 %
	Alemán	376.65	2.60 %	0.18 %
MultiLing	Inglés	1979.46	4.36 %	1.27 %
	Español	3054.90	6.38 %	1.76 %
	Alemán	1999.76	3.84 %	0.43 %

Tabla 3: Valores medios de las estadísticas de los documentos. PPD: Palabras por documento (sin stopwords)

Como se puede apreciar, no existe una técnica concreta que represente los mejores resultados para los tres idiomas, por lo que es difícil generalizar y determinar el mejor enfoque. No obstante, se destaca el enfoque sin análisis semántico (*PCA_base*) ya que presenta muy buenos resultados, siendo muy interesante dado que es totalmente independiente del idioma, al contrario que los métodos con conocimiento léxico-semántico, en los que hay que disponer de sistemas de reconocimiento de entidades y recursos para identificar conceptos específicos para cada idioma, cuya repercusión y correcto funcionamiento condicionan en gran medida los resultados obtenidos.

Por otro lado, los resultados (ROUGE-1) para el idioma español son interesantes ya que ofrece su mejor valor con H2 cuando incorporamos co-

nocimiento léxico-semántico para ambos corpus, obteniendo los mejores resultados en comparación con el resto de idiomas.

En general, la aportación de conocimiento léxico-semántico en ambos corpus es mínima, dando lugar a que su contribución no sea excesivamente notable. En el caso del corpus MultiLing 2015 aunque en porcentaje la aportación de NER es mayor que para el corpus JRC, ese porcentaje tiene menor efecto dada la extensión total de los documentos a resumir. Es por ello que el enfoque que incluye NER mejora en algunos casos para el corpus JRC (sobre todo para las heurísticas H1 y H2), pero no para el corpus MultiLing 2015.

En comparación con el enfoque independiente del idioma, las mejorar con la incorporación de conocimiento léxico-semántico no son las esperadas. Esto puede deberse, en parte, a que el uso de recursos y herramientas externas para identificar tanto NERs como conceptos sinónimos pueda dar lugar a la presencia de errores cometidos por las propias herramientas, o factores como la no correcta asignación de un sentido a una palabra, ya que se realiza el tipo de desambiguación más básica. Por ello, de los resultados obtenidos, así como del análisis en detalle de algunos resumenes generados se han identificado una serie de limitaciones no contempladas inicialmente en nuestro enfoque, y que podrían afectar negativamente a la calidad de los resúmenes. La primera de ellas es el uso de conocimiento semántico sin ninguna técnica de desambiguación. Debido a que la tarea de desambiguación es muy compleja y que no era el objetivo principal de este trabajo, optamos por realizar la identificación de conceptos según su sentido más frecuente, y por lo tanto, no teniendo en cuenta el contexto en el que se está utilizando el término en cuestión. A pesar de que los resultados para el sentido más frecuente giran en torno al 50 % (McCarthy, 2011), esto puede dar lugar a que: i) no se estén identificando correctamente algunos conceptos; y ii) se cometan errores en la agrupación de los conceptos. Técnicas más recientes y precisas en la tarea de desambiguación (Agirre et al., 2014), así como el uso de recursos más actualizados y con mayor cobertura, como Multilingual Central Repository o GermaNet podrían contribuir a mejorar los resultados.

Por otro lado, una vez calculada la matriz obtenida a partir de la técnica PCA, estamos teniendo en cuenta todas las palabras que integran cada uno de los conceptos identificados para realizar la selección de las frases hasta que se alcanza una determinada longitud. Hubiera sido interesante analizar y aplicar alguna técnica para realizar una segunda selección de entre todas esas frases para que los resúmenes generados recogieran una mayor variedad de conceptos, ya que debido a la longitud impuesta por los resúmenes modelo, éstos se generaron con el tamaño especificado.

Tal y como se comenta en la sección 8, se plantea abordar estos y otros aspectos en los trabajos futuros para determinar si solventando estas limitaciones, la aportación de conocimiento léxico-semántico tiene una influencia positiva mayor en la técnica PCA o, si por el contrario, la influencia es ngativa.

7 Comparativa con respecto a sistemas existentes

Una vez analizados los resultados, en esta sección vamos a realizar una comparativa de los mejores resultados de nuestros métodos con respecto a algunos sistemas de resúmenes multilingües existentes. Los sistemas utilizados se han seleccionado dada su disponibilidad y accesibilidad para la generación de resúmenes multilingües con ambos corpus. En concreto, dichos sistemas son:

- Open Text Summarizer (OTS)[10]. El enfoque implementado en este sistema identifica las palabras clave mediante la ocurrencia de las palabras. Emplea algunos recursos específicos por idioma tales como analizadores lingüísticos y listas de stopwords para más de 25 idiomas.
- Resumidor integrado en Microsoft Word 2007 (MS Word)[11]. Dado que es un sistema comercial, los detalles de implementación no son públicos.
- Essential Summarizer (Essential)[12]. Este sistema es una versión comercial del presentado por (Lehmam, 2010). Se basa en técnicas lingüísticas para realizar análisis semántico, teniendo en cuenta los elementos discursivos del texto.

Además, dado que el corpus JRC fue previamente utilizado por (Lloret & Palomar, 2011), se incluye en la comparativa sus mejores resultados obtenidos por el enfoque dependiente del idioma (LS), que emplea recursos específicos para cada idioma (etiquetador gramatical, NER e identificación de conceptos), siendo similar a nuestro método (c) *PCA_base+CPT+NER*. La Tabla 4 muestra la comparativa realizada.

[10] http://libots.sourceforge.net/

[11] https://support.office.com/en-nz/article/Automatically-summarize-a-document-b43f20ae-ec4b-41cc-b40a-753eed6d7424

[12] https://essential-mining.com/

	Sistema	Inglés	Español	Alemán
JRC	Mejor (a)	**0.57797**	0.58786	0.53300
	Mejor (b)	0.54576	0.59845	0.53364
	Mejor (c)	0.56614	0.59878	**0.54070**
	LS	0.56530	**0.62351**	0.52614
	OTS	0.55732	0.60591	0.53451
	MS Word	0.53591	0.57396	0.48427
	Essential	0.52622	0.53978	0.43727
MultiLing	Mejor (a)	**0.43991**	0.44031	0.26939
	Mejor (b)	0.43785	0.44024	**0.27202**
	Mejor (c)	0.43633	**0.44125**	0.26873
	OTS	0.43090	0.41345	0.26293
	MS Word	0.43382	0.40501	0.27096
	Essential	0.41382	0.39131	0.24127

Tabla 4: Comparativa (R-1, medida F) con diferentes enfoques - (a)*PCA_base*; (b)*PCA_base+CPT*; (c)*PCA_base+CPT+NER*.

Como se puede observar, nuestros enfoques presentan los mejores resultados para el corpus de entrenamiento MultiLing 2015, superándolos para todos los idiomas. Con respecto al corpus JRC, nuestros enfoques mejoran los resultados para alemán e inglés pero no consiguen superar a los de OTS y LS para el idioma español, a pesar de quedarse muy cercanos. La razón por la que nuestro método no haya sido capaz de obtener mejores resultados que el método LS de (Lloret & Palomar, 2011) puede deberse a que en nuestro enfoque no se utiliza ningún método para la desambiguación del sentido de las palabras, mientras que en el trabajo de referencia analizan los documentos en español mediante el analizador Freeling (Padró & Stanilovsky, 2012), que procesa y anota semánticamente un texto utilizando algoritmos de desambiguación como UKB (Agirre & Soroa, 2009), que han demostrado superar a la aproximación del sentido más frecuente. Cabe destacar el enfoque (c) ya que la adición de conocimiento léxico-semántico se hace de manera similar a LS, consiguiéndose mejorar los resultados para el inglés y alemán. El enfoque base (a) presenta buenos resultados y puede ser generalizable para cualquier idioma dado que su ejecución es invariante del idioma. Por lo tanto, se puede concluir que los métodos propuestos presentan resultados muy competitivos comparados con sistemas comerciales.

8 Conclusiones y trabajos futuros

En este artículo se ha realizado un estudio de la técnica PCA para la generación de resúmenes extractivos mono-documento y multilingües, analizando la influencia de introducir conocimiento léxico-semántico a la técnica base (reconocimiento de entidades e identificación de conceptos sinónimos) e investigando cuatro heurísticas diferentes para seleccionar la frase a partir de las palabras clave determinadas con la técnica PCA.

Para la experimentación se utilizaron dos corpus de diferente naturaleza (noticias periodísticas y artículos de la Wikipedia) y se generaron resúmenes automáticos para tres idiomas (inglés, español y alemán), evaluando la relevancia de la información seleccionada con respecto a resúmenes modelos utilizando la herramienta ROUGE.

Como conclusión general, la calidad de los resúmenes con conocimiento léxico-semántico es muy dependiente de las características de los textos que hay que resumir dado que la cantidad de entidades y conceptos que posean afecta en gran medida a los resultados. Los resúmenes generados con la técnica sin ningún tipo de conocimiento (*PCA_base*) presentan muy buenos resultados en comparación con otros sistemas existentes, siendo un enfoque atractivo dada su independencia del idioma con que se trabaje.

Para futuros trabajos, se propone mejorar el enfoque propuesto incluyendo el etiquetado gramatical, que podría mejorar notablemente el rendimiento de la etapa de identificación de conceptos, y replicar la experimentación utilizando recursos léxico-semánticos más actualizados, como el recurso Multilingual Central Repository para el castellano y el recurso GermaNet para el alemán. También planteamos redefinir la estrategia de selección de palabras clave, para realizar un filtrado intermedio de palabras clave relevantes y evitar así considerar todas las palabras de la matriz obtenida con la técnica PCA. Finalmente, sería interesante analizar los textos de los corpus para poder realizar y orientar el resumen que más se pueda adecuar, dado que cada heurística da lugar a un tipo de resumen diferente.

Agradecimentos

Esta investigación se ha realizado gracias a la financiación recibida en los proyectos: DIIM2.0: Desarrollo de técnicas Inteligentes e Interactivas de Minería y generación de información sobre la web 2.0 (PROMETEOII/2014/001) de la Generalitat Valenciana; SAM (FP7-611312) de la Comisión Europea; "Análisis de Tendencias Mediante Técnicas de Opinión Semántica" (TIN2012-38536-C03-03) y "Técnicas de Deconstrucción en la Tecnologías del Lenguaje Humano" (TIN2012-31224)), del Ministerio de Economía y Competitividad del Gobierno de España; "Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario" (GRE13-15), de la Universidad de Alicante.

Referencias

Agirre, Eneko, Oier López de Lacalle & Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Comput. Linguist.* 40(1). 57–84.

Agirre, Eneko & Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. En *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* EACL '09, 33–41. Stroudsburg, PA, USA: Association for Computational Linguistics.

Barzilay, Regina & Michael Elhadad. 1999. Using lexical chains for text summarization. En *Inderjeet Mani and Mark Maybury, editors,* Advances in Automatic Text Summarization, 111–122. MIT Press.

Conroy, John M, Sashka T Davis, Jeff Kubina, Yi-Kai Liu, Dianne P. O'Leary & Judith D. Schlesinger. 2013. Multilingual Summarization: Dimensionality Reduction and a Step Towards Optimal Term Coverage. En *MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, 55–63.

Cowie, Jim, Kavi Mahesh, Sergei Nirenburg & Remi Zajac. 1998. MINDS - Multi-lingual INteractive Document Summarization. *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization* 122–123.

Dunteman, George H. 1989. *Principal components analysis* 69. Sage.

Estellés Arolas, Enrique, Fernando González Ladrón De Guevara & Antonio Falcó Montesinos. 2010. Principal Component Analysis for Automatic Tag Suggestion. Relatório técnico. http://dspace.ceu.es/bitstream/10637/6327/1/Principal%20component%20analysis%20for%20automatic%20tag%20suggestion.pdf.

Giannakopoulos, G, M El-Haj, J Steinberger, B Favre, M Litvak & V Varma. 2011. TAC 2011 MultiLing Pilot Overview. *TAC 2011 Workshop* .

Giannakopoulos, George. 2013. Multi-document multilingual summarization and evaluation tracks in acl 2013 multiling workshop. En *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, 20–28. Sofia, Bulgaria: Association for Computational Linguistics.

Gupta, Vishal & Gurpreet Singh Lehal. 2010. A Survey of Text Summarization Extractive Techniques. *Journal of Emerging Technologies in Web Intelligence* 2(3). 258–268.

Hovy, Eduard & Chin-yew Lin. 1997. Automated Text Summarization in SUMMARIST. En *ACL Workshop on Intelligent, Scalable Text Summarization*, 18–24.

Kubina, Jeff, John M Conroy & Judith D Schlesinger. 2013. ACL 2013 MultiLing Pilot Overview. En *MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, 29–38.

Lee, Chang Beom, Min Soo Kim & Hyuk Ro Park. 2003. Automatic Summarization Based on Principal Component Analysis. *Progress in Artificial Intelligence* 409–413.

Lehmam, Abderrafih. 2010. Essential summarizer: innovative automatic text summarization software in twenty languages. En *RIAO '10: Adaptivity, personalization and fusion of heterogeneous information*, 216–217. Le Centre de Hautes Etudes Internationales D'Informatique Documentaire.

Lin, Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries. En *Marie-Francine Moens, S. S., editor, Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 74–81.

Litvak, Marina & Mark Last. 2013. Multilingual Single-Document Summarization with MUSE. En *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, 77–81. Association for Computational Linguistics.

Lloret, Elena & Manuel Palomar. 2011. Finding the Best Approach for Multi-lingual Text Summarisation: A Comparative Analysis. En *International Conference Recent Advances in Natural Language Processing* Sep., 194–201.

Mani, Inderjeet & Mark T. Maybury. 1999. *Advances in automatic text summarization.* The MIT Press. ISBN 0-262-13359-8.

McCargar, Victoria. 2005. Statistical approaches to automatic text summarization. *Bulletin of the American Society for Information Science and Technology* 30(4). 21–25.

McCarthy, Diana. 2011. Word sense disambiguation. Seminar.

Miller, George A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38(11). 39–41.

Nenkova, Ani & Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval* 5(2-3). 103–233.

Padró, Lluís & Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. En *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey: ELRA.

Patel, Alkesh, Tanveer Siddiqui & U. S. Tiwary. 2007. A language independent approach to multilingual text summarization. *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)* .

Ringnér, Markus. 2008. What is principal component analysis? *Nature biotechnology* 26(3). 303–304.

Sparck-Jones, Karen. 1999. Automatic summarising: factors and directions. *Advances in Automatic Text Summarization* 1–21.

Spärck Jones, Karen. 2007. Automatic summarising: The State of the Art. *Information Processing & Management* 43(6). 1449–1481.

Steinberger, Josef. 2013. The uwb summariser at multiling-2013. En *Proceedings of the MultiLing 2013 Workshop on Multilingual Multidocument Summarization*, 50–54. Sofia, Bulgaria: Association for Computational Linguistics.

Teng, Zhi, Ye Liu, Fuji Ren, Seiji Tsuchiya & Fuji Ren. 2008. Single document summarization based on local topic identification and word frequency. En *Proceedings of the Seventh Mexican International Conference on Artificial Intelligence*, 37–41. Washington, DC, USA: IEEE Computer Society.

Tjong, Erik F, Kim Sang & Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. En *7th conference on Natural language learning at HLT-NAACL 2003*, vol. 4, 142–147. Association for Computational Linguistics.

Uzêda, Vinícius Rodrigues, Thiago Alexandre Salgueiro Pardo & Maria Das Graças Volpe Nunes. 2010. A comprehensive comparative evaluation of rst-based summarization methods. *ACM Trans. Speech Lang. Process.* 6(4). 4:1–4:20.

Vossen, Piek. 2004. Eurowordnet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual index. *International Journal of Lexicography Vol.17* 2. 161–173.

Projetos, Apresentam-Se!

A arquitetura de um glossário terminológico Inglês-Português na área de Eletrotécnica

The architecture of an English-Portuguese glossary in the field of Electrical Terminology

Sabrina Bonqueves Fadanelli
Universidade Federal do Rio Grande do Sul (UFRGS)
sabrina_mina2006@hotmail.com

Maria José Bocorny Finatto
Universidade Federal do Rio Grande do Sul (UFRGS)
mariafinatto@gmail.com

Resumo

Neste artigo descrevemos alguns dos procedimentos para a execução de um protótipo online de glossário Inglês-Português na área de Eletrotécnica/Engenharia Elétrica – voltado principalmente aos alunos iniciantes dos cursos técnicos de Eletrotécnica e de graduação em Engenharia Elétrica/Eletrotécnica. A metodologia envolveu um *corpus* de *datasheets*, documentos muito utilizados por profissionais das áreas de Elétrica e Eletrotécnica; e a comparação com os dados obtidos de 108 alunos iniciantes das áreas Elétricas. Os resultados apontam para a relevância de se considerar o olhar do público-alvo para a compilação do glossário.

Palavras chave

Terminografia, glossário, datasheet, Eletrotécnica

Abstract

This article describes some of the procedures for the execution of an online English-Portuguese glossary prototype in Eletrical Engineering / Eletrotechnical Field terminology – aimed mainly at beginner students from technical and graduation courses in Electrical Engineering. The methodology is comprised of a corpus of *datasheets*, documents often used by professionals of the Electrical Engineering area, and the comparison of data obtained from these *datasheets* with the data gathered from 108 students of Electrical courses. Results point to the relevance of considering the point of view of our target audience to build the glossary properly.

Keywords

Terminography, pedagogical, glossary, datasheet, eletrotechnical field

1 Introdução

Até que se tornem especialistas ou proficientes na terminologia de sua área, estudantes de Eletrotécnica e Engenharia Elétrica no Brasil devem aprimorar seus conhecimentos em Língua Inglesa, já que praticamente toda a documentação de componentes elétricos se encontra nesta língua. A Eletrotécnica é uma área da Engenharia Elétrica que lida com instalações de redes elétricas, transformadores, circuitos elétricos residenciais, industriais, etc. Seu material principal de consulta são os *datasheets*, documentos que contêm em sua estrutura informações técnicas sobre os dispositivos e aparelhos elétricos. Muitos dos estudantes iniciantes nestas áreas encontram dificuldades na leitura destes documentos técnicos, já que muitas vezes não possuem um conhecimento proficiente da Língua Inglesa, nem da área técnica em si. Isso demonstra uma necessidade de se produzir uma ferramenta que forneça tanto a tradução dos termos de Eletrotécnica em português, como também apresente definições explicativas e didáticas voltadas aos alunos iniciantes da Eletrotécnica/Engenharia Elétrica ou a possíveis usuários que apresentem necessidade de utilização da ferramenta. Além disso, a ferramenta deverá conter exercícios que possam auxiliar na leitura e compreensão dos *datasheets*. O desenvolvimento desta ferramenta, assim sendo, deverá seguir parâmetros pertencentes a uma terminografia que atenda às necessidades deste público-alvo.

O objetivo deste artigo é descrever alguns procedimentos usados para a seleção de termos que venham a integrar um protótipo de glossário online Inglês- Português na área de Eletrotécnica/Engenharia Elétrica, sob a ótica da

Terminologia Textual.

A perspectiva da Terminologia Textual considera o texto especializado como o habitat natural das terminologias, sendo ele o todo que vai determinar seus modos de dizer específicos (Kilian, 2007; Zilio, 2010). Com base nisso, não só a terminologia em si é focada, mas também outros aspectos textuais como macro e microestruturas do texto: as frases, os sintagmas, os gêneros textuais, etc. (Finatto, 2004, 2010). Como afirmam Bourugault & Slodzian (2004, pg. 107), a abordagem textual permite visualizar todas as "experiências" da análise linguística textual, não considerando os termos como "unidades de conhecimento que habitam a língua".

2 Os *datasheets*

Os documentos técnicos chamados de *datasheets* fornecem informações como potência, resistência, dimensões, etc., não somente com a finalidade de compreender o funcionamento de um componente, mas também para não danificá-lo e garantir a segurança do usuário (ver Figura 1 para exemplos de *datahseets*). De acordo com Dewey (1998), o *datasheet* é o conector comunicativo entre indivíduos ou empresas com necessidades diferentes e níveis de conhecimento diferentes. A relação entre os engenheiros que escrevem os *datasheets* e os seus possíveis leitores pode variar: comprador do produto especificado pelo *datasheet* x vendedor; equipe de produção e engenharia do produto, pessoal do *marketing*, e inspetores de qualidade, além de outros.

3 A Metodologia

Nesta investigação, escolhemos a metodologia da Linguística de Corpus como apoio. A Linguística de *Corpus* faz uso da compilação e análise de elementos e estruturas linguísticas de um determinado *corpus* por meio de sistemas computadorizados (Berber-Sardinha, 2004; Biber, 1988). A utilização da Linguística de *Corpus* nos permitirá analisar a ocorrência de termos nos *datasheets*.

Além disso, levamos em consideração não somente o material de onde os termos serão retirados, mas também o ponto de vista das pessoas que o utilizarão: os alunos, fazendo uma comparação entre os dados obtidos de ambas as fontes.

A obtenção de dados de extratores foi realizada utilizando duas ferramentas, o AntConc (Anthony, 2004) e o TermoStat (Drouin, 2003). Estas foram escolhidas por serem de natureza diferente, uma estatística e a outra linguística, a

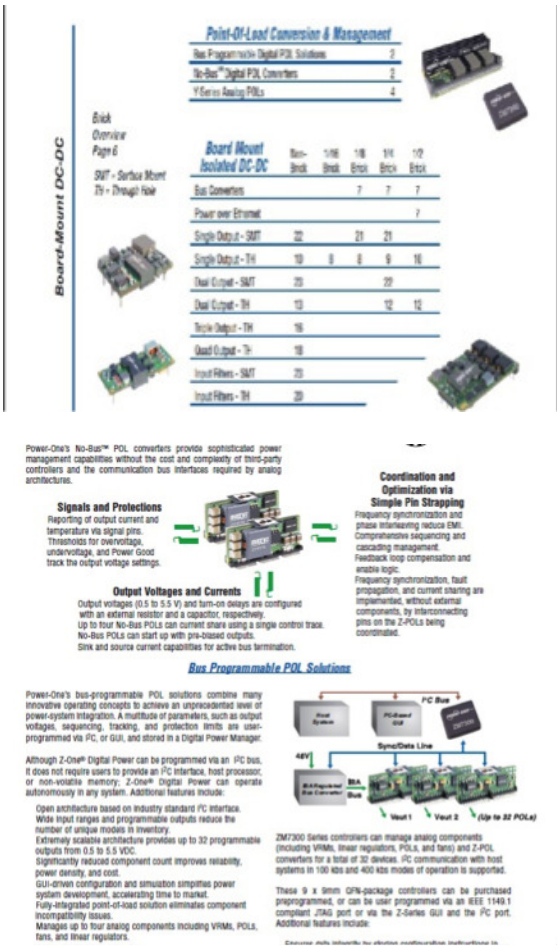


Figura 1: Exemplos de *datasheets*.

fim de promover mais variedade nos resultados (Vieira & Lopes, 2010).

O corpus de estudo foi coletado a partir de *datasheets* em Língua Inglesa de componentes elétricos e áreas da Eletrotécnica sugeridas por dois Engenheiros Elétricos como essenciais para o conhecimento técnico dos aprendizes. Os *datasheets* foram separados em 11 (onze) componentes/áreas pertinentes à Eletrotécnica e transformados em 11 (onze) arquivos em formato txt.

No AntConc, foi utilizada a função *Keyword* — que extrai as palavras- chave de um *corpus*, comparando o *corpus* de estudo com um outro *corpus* — chamado de *corpus* de referência; o *corpus* de referência foi composto de textos extraídos do *corpus Coca*[1] (*Contemporary Corpus of American English*) e do *BNC*[2] (*British National Corpus*), aleatoriamente. O *TermoStat* (Drouin, 2003) é um extrator de termos acessado gratuitamente na web, que junta candidatos a termos (doravante tratados como CTs) de acordo

[1] O *Corpus of Contemporary American English* está disponível em `http://corpus.byu.edu/coca/`.

[2] O *British National Corpus* está disponível em `http://www.natcorp.ox.ac.uk/`.

com padrões gramaticais e frequência.

Como próximo passo, exatamente os mesmos *datasheets* utilizados na extração automática foram distribuídos a 108 alunos de dois cursos: Técnico em Eletrotécnica da FTEC Faculdades e alunos no primeiro semestre de Engenharia Elétrica da Universidade de Caxias do Sul (UCS). Foi-lhes instruído que marcassem com um círculo ou sublinhassem as palavras dos *datasheets* que eles não compreendiam durante uma leitura cuidadosa do documento. As comparações entre os dados apresentados pelos extratores e pelos alunos levaram em conta as seguintes questões:

A) Quantos CTs em comum/exclusivos foram apresentados entre os extratores e os dados dos alunos?

B) Os CTs apontados pelos alunos seriam em sua maior proporção vocabulário pertencente à área de Eletrotécnica/Engenharia Elétrica?

C) Os CTs que aparecem mais frequentemente nos extratores são os mesmos que os apontados pelos alunos?

3.1 A) Quantos CTs em comum/exclusivos foram apresentados entre os extratores e os dados dos alunos?

Para que o Microsoft Excel separasse os candidatos em comum entre os extratores e alunos, utilizou-se a função *PROCV*. A Tabela 1 permite uma visualização dos totais de termos extraídos do AntConc e TermoStat e apontados pelos alunos. Podemos ver que a quantidade de termos em comum foi baixa.

Total de termos TermoStat	782
Total de termos AntConc	533
Total de termos alunos	640
Termos em comum TermoStat + Alunos	73
Termos em comum AntConc + Alunos	40
Termos em comum AntConc + TermoStat	134

Tabela 1: Termos em comum.

De certa forma este resultado já era esperado, justamente pela natureza diferente entre as duas ferramentas (uma de abordagem estatística e outra de abordagem linguística) e com relação aos alunos, que se utilizam de estratégias cognitivas para a escolha dos termos que desconhecem. A baixa proporção de termos em comum indica a utilidade de se combinar dados dos extratores com dados apontados pelo público-alvo, já que se utilizarmos somente os extratores poderemos não incluir informações importantes para os usuários.

3.2 B) Os CTs apontados pelos alunos seriam em sua maior proporção vocabulário pertencente à área de Eletrotécnica/Engenharia Elétrica?

Um dos questionamentos que se fez presente quando da tentativa de selecionar os termos mais relevantes para os usuários do protótipo de glossário é se os alunos apresentam mais dificuldade em interpretar/compreender termos que pertençam à área de Eletrotécnica, ou, em outras palavras, que sejam conceitos especializados. Desse modo, a fim de se determinar a quantidade de CTs apontados pelos alunos que também possam ser considerados conceitos especializados, decidiu-se por utilizar uma ferramenta online que situe o vocabulário no contexto cultural e socioprofissional da Eletrotécnica/Engenharia Elétrica: a Electropedia[3]. A Electropedia é um banco de dados terminológicos organizado pela IEC – International Electrotechnical Comission (Comissão Internacional de Eletrotécnica), a organização mundial que prepara e padroniza todas as configurações e características de aparelhos e dispositivos elétricos, voltada em sua essência para especialistas da área, ou ao menos a alguém que possua uma certa fluência em Língua Inglesa.

Manualmente, os CTs em comum entre AntConc, TermoStat e alunos e os CTs exclusivos dos alunos foram inseridos na Electropedia. Não foram inseridos os CTs exclusivos dos extratores pois, o objetivo aqui era verificar o vocabulário apontado pelos alunos. Os CTs que não foram encontrados na Electropedia passaram por uma avaliação de um engenheiro elétrico. A comparação entre os resultados obteve o seguinte: do total de 640 CTs apontados pelos alunos, 295 foram encontrados e 345 não foram encontrados na Electropedia. Destes 345 não encontrados, cerca de 101 eram fórmulas, abreviações e termos que o especialista considerou técnicos, mesmo não aparecendo na Electropédia. Sobraram então 244 CTs que nem o engenheiro considerou como vocabulário técnico, nem foram encontrados na Electropedia. Isso representa cerca de 38% dos CTs apontados pelos alunos.

Embora os dados obtidos respondam à questão de forma a confirmar a predominância de vocabulário pertencente à área de Eletrotécnica/Engenharia Elétrica, os resultados anteriormente descritos mostram que os alunos apresentaram dúvidas em uma porcentagem relevante de CTs não considerados conceitos especializados; conclui-se, então, que incluir estes

[3]Disponível em Disponível em: http://www.electropedia.org/

últimos na compilação do glossário, seja nas definições, seja nos exercícios, provavelmente auxiliaria os usuários em sua leitura dos *datasheets*.

3.3 C) Os CTs que aparecem mais frequentemente nos extratores são os mesmos que os apontados pelos alunos?

Neste estudo, perguntamos se a frequência dos termos já apontados como chave pelos extratores seria um fator que refletiria informações relevantes para a compilação de nosso glossário. Assim sendo, comparamos se os CTs apontados pelos alunos coincidiam em sua maioria com os CTs considerados mais frequentes nos extratores.

Primeiramente fez-se necessário determinar quais CTs nos extratores eram considerados frequentes. Com base na normatização proposta por Biber et al. (1998) a fim de estabelecer aproximadamente quantas vezes uma palavra apareceria em cada mil palavras de um corpus, representando ao menos 1% da frequência (ou dez ocorrências em cada mil), o seguinte cálculo foi usado: quantidade de vezes que um termo aparece, dividido pelo numero de termos totais, multiplicado por mil. Todos os 533 CTs resultantes do AntConc já representavam cerca de 1% de frequência, pois a função *keyness* apenas aponta termos que aparecem seis vezes ou mais (6 ÷ 533 x 1000 = + ou – 11). No TermoStat, determinou-se que 1% de frequência significava 8 ocorrências (8 ÷ 782 x 1000 = 10.23).

Os CTS foram colocados na planilha *Excel* e novamente a fórmula *PROCV* foi aplicada. A Tabela 2 mostra quantos CTs apontados pelos alunos estão entre os CTs dos extratores considerados mais frequentes.

	AntConc	**TermoStat**
Total	533	782
Total CTs 1% ou + de frequência	533	64
Total CTs alunos entre os + frequentes dos extratores	142	14

Tabela 2: Comparação de termos frequentes.

Podemos ver que considerar a frequência como um fator determinante na escolha de CTs para o glossário não é muito recomendável. A maioria dos termos que causaram mais dificuldade aos alunos não estão entre os listados como os mais frequentes pelos extratores. Assim sendo, se utilizássemos somente os dados fornecidos pelos extratores estaríamos excluindo dados importantes para a confecção do glossário visando a necessidade dos alunos.

4 Outras Observações

Uma análise mais cuidadosa com os CTs apontados pelos alunos pareceu mostrar que a habilidade de considerar itens periféricos do texto (como combinações de palavras, elementos próximos a figuras, gráficos, reconhecimento de cognatos, etc.) para interpretar um termo aparentemente desconhecido não se fez presente em alguns termos apontados pelos alunos. A análise foi feita examinando cada um dos 640 CTs resultantes da coleta de dados com os aprendizes de Eletrotécnica/Engenharia Elétrica, em seu contexto de ocorrência, com a ferramenta AntConc. Dentre estes cerca de 84 CTs eram cognatos, e muitos dos que não eram cognatos poderiam ter sido compreendidos se levados em conta juntamente com as palavras ou outros elementos que os acompanhavam. Essa característica fornece mais um possível critério para a compilação do glossário: os termos, quando transformados em verbetes no glossário, não devem estar sozinhos — ou seja, cada verbete deve vir acompanhado de seus colocados como apareceram no corpus. E os exercícios devem levar em conta cognatos e as combinatórias lexicais que acompanham cada CT.

Outra questão de grande relevância na análise foram os termos encontrados em tabelas, gráficos e figuras, elementos muito presentes no gênero textual *datasheet*. Cerca de 46% dos termos contidos nestes elementos coincidem com os termos que foram apontados pelos alunos, uma quantidade bastante considerável que indica a importância de serem incluídos no glossário.

4.1 Conclusões

Este estudo nos forneceu dados que permitem afirmar que o desenvolvimento do glossário a partir da consideração do ponto de vista do público-alvo aumenta as chances da ferramenta ser considerada útil por aprendizes ainda não especialistas, pois resulta em uma seleção de termos que vai ao encontro não somente de conceitos especializados que os alunos precisam para ler os *datasheets*, mas também às outras palavras que compõem a tessitura do documento e que possam ser desconhecidas para os alunos.

Referências

Anthony, Laurence. 2004. AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. Em *Proceedings of IWLeL: An*

interactive workshop on language e-learning, 7–13.

Berber-Sardinha, Tony. 2004. *Linguística de corpus*. São Paulo: Manole.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Bourugault, Didier & Monique Slodzian. 2004. Por uma terminologia textual. Em Maria da Graça Krieger & Luzia Araújo (eds.), *A terminologia em foco. cadernos de tradução* 17, Porto Alegre: Instituto de Letras da UFRGS.

Dewey, F. Raymond. 1998. A complete guide to datasheets. *Sensors Magazine* Disponível em http://www.allegromicro.com/~/media/Files/Technical-Documents/pub26000-Complete-Guide-To-Datasheets.ashx.

Drouin, Patrick. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology* 9(1). 99–117.

Finatto, Maria José Bocorny. 2004. Termos, textos e textos com termos: novos enfoques dos estudos terminológicos de perspectiva linguística. Em Aparecida Negri Isquerdo & Maria Graça Krieger (eds.), *As ciências do léxico: lexicologia, lexicografia, terminologia*, vol. 2, 353–363. Campo Grande: UFMS/PPG-Letras UFRGS.

Finatto, Maria José Bocorny. 2010. Estudos sobre linguagens e textos científicos e técnicos: o que é uma terminologia textual? Em Gisela Collischon (ed.), *Encontro do CELSUL*, 153–172. UNISUL.

Kilian, Cristiane. 2007. *A retomada de unidades de significação especializada em textos em língua alemã e portuguesa sobre gestão de resíduos: uma contribuição para a tradução técnico científica*. Porto Alegre: Universidade Federal do Rio Grande do Sul. Tese de Doutoramento.

Vieira, Renata & Lucelene Lopes. 2010. Processamento de linguagem natural e o tratamento computacional de linguagens científicas. Em Cristina Becker Lopes Perna, Heloísa Orsi Koch Delgado & Maria José Bocorny Finatto (eds.), *Linguagens especializadas em corpora: modos de dizer e interfaces de pesquisa*, Porto Alegre: EDIPUCRS.

Zilio, Leonardo. 2010. Terminologia textual e linguística de corpus: estudo em parceria. Em Cristina Becker Lopes Perna, Heloísa Orsi Koch Delgado & Maria José Bocorny Finatto (eds.), *Linguagens especializadas em corpora: modos de dizer e interfaces de pesquisa*, Porto Alegre: EDIPUCRS.

http://www.linguamatica.com/

linguaMATICA