Volume 5, Número 1- Julho 2013

# linguaMATICA

ISSN: 1647-0818

UNIVERSIDADE DE VIGO

Universidade do Minho

Volume 5, Número 1 – Julho 2013

# LinguaMÁTICA

ISSN: 1647–0818

# Conteúdo

**I Dossier** **11**

**Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual Central Repository**
*Aitor Gonzalez-Agirre e German Rigau* . . . . . 13

**II Artigos de Investigação** **29**

**Un método de análisis de lenguaje tipo SMS para el castellano**
*Andrés Alfonso Caurcel Díaz, José María Gómez Hidalgo e Yovan Iñiguez del Rio* 31

**Extracção de relações semânticas de textos em português explorando a DBpédia e a Wikipédia**
*David S. Batista, David Forte, Rui Silva, Bruno Martins e Mário J. Silva* . . . 41

**Clasificación automática del registro lingüístico en textos del español: un análisis contrastivo**
*John A. Roberto, Maria Salamó e M. Antònia Martí* . . . . . 59

# Editorial

*Benvolguts i benvolgudes:*

*amb la publicació d'aquest volum 5.1 de Linguamática, assolim els deu exemplars d'aquesta revista i entrem amb il·lusió i forces renovades amb el seu cinquè any d'edició. Han sigut cinc anys difícils, en els quals hem hagut de treballar amb molt d'esforç per aconseguir oferir-vos semestre a semestre una revista científica de qualitat sobre el processament de les llengües ibèriques, sense renunciar a l'objectiu de fer ciència en les nostres llengües, per a les nostres comunitats lingüístiques i fent servir els nostres idiomes com a mitjà natural i privilegiat de comunicació.*

*A partir d'aquest número de Linguamática, i per tal de facilitar encara més la indexació de la revista (un assumpte especialment dificultós a hores d'ara per a les publicacions en llengües diferents de l'anglesa), els articles publicats portaran, a més d'un resum i unes paraules clau en anglès, la traducció del títol de l'article en aquesta llengua. Amb aquesta nova metadada, pretenem facilitar la feina d'indexació de la revista en índexs internacionals i garantir una difusió encara més gran dels articles publicats a Linguamática en qualsevol llengua de la nostra península.*

*El nostre agraïment més sincer per a tots els col·laboradors i lectors de Linguamática, sense els quals no tindria cap sentit seguir tirant endavant aquest projecte.*

*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**,
ISLA e Universidade Nova de Lisboa

**Anselmo Peñas**,
Univers. Nac. de Educación a Distancia

**Antón Santamarina**,
Universidade de Santiago de Compostela

**Antonio Moreno Sandoval**,
Universidad Autónoma de Madrid

**António Teixeira**,
Universidade de Aveiro

**Arantza Díaz de Ilarraza**,
Euskal Herriko Unibertsitatea

**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**,
Universidade Beira Interior

**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

**Joaquim Llisterri**,
Universitat Autònoma de Barcelona

**José Carlos Medeiros**,
Porto Editora

**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

**Lluís Padró**,
Universitat Politècnica de Catalunya

**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

**Patrícia Cunha França**,
Universidade do Minho

**Pablo Gamallo Otero**,
Universidade de Santiago de Compostela

**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

# Dossier

# Construcción de una base de conocimiento léxico multilíngüe de amplia cobertura: Multilingual Central Repository

**Building a wide coverage multilingual lexical knowledge base: Multilingual Central Repository**

Aitor Gonzalez-Agirre
Universidad del País Vasco
aitor.gonzalez-agirre@ehu.es

German Rigau
Universidad del País Vasco
german.rigau@ehu.es

**Resumen**

El uso de recursos semánticos de amplia cobertura y dominio general se ha convertido en una práctica común y a menudo necesaria para los sistemas actuales de Procesamiento del Lenguaje Natural (PLN). WordNet es, con mucho, el recurso semántico más utilizado en PLN. Siguiendo el éxito de WordNet, el proyecto EuroWordNet ha diseñado una infraestructura semántica multilingüe para desarrollar wordnets para un conjunto de lenguas europeas. En EuroWordNet, estos wordnets están interconectados con enlaces interlingüísticos almacenados en el índice interlingual (en inglés, *interlingual-index* o ILI). Siguiendo la arquitectura de EuroWordNet, el proyecto MEANING ha desarrollado las primeras versiones del *Multilingual Central Repository* (MCR) usando un ILI basado en WordNet 1.6. Con ello, se mantiene la compatibilidad entre los wordnets de diferentes idiomas y versiones. Esta versión del MCR integra seis versiones diferentes de la WordNet inglés (de 1.6 a 3.0) y también wordnets en castellano, catalán, euskera e italiano, junto a más de un millón de relaciones semánticas entre conceptos así como propiedades semánticas de diferentes ontologías. Recientemente hemos desarrollado una nueva versión del MCR usando un ILI basado en WordNet 3.0. Esta nueva versión del MCR integra wordnets de cinco idiomas diferentes: inglés, castellano, catalán, euskera y gallego. La versión actual del MCR, al igual que la anterior, integra sistemáticamente miles de relaciones semánticas entre conceptos. Además, el MCR se ha enriquecido con cerca de 460.000 propiedades semánticas y ontológicas que incluyen *Base Level Concepts*, *Top Ontology*, *WordNet Domains* y *AdimenSUMO*, proporcionando coherencia ontológica a todos los wordnets y recursos semánticos integrados en ella.

**Palabras clave**

Semántica Léxica, Bases de Conocimiento Léxico, WordNet, EuroWordNet

**Abstract**

The use of wide coverage and general domain semantic resources has become a common practice and often necesary by existing systems Natural Language Processing (NLP). WordNet is by far the most widely used semantic resource in NLP. Following the success of WordNet, the EuroWordNet project has designed a multilingual semantic infrastructure to develop wordnets for a set of European languages. In EuroWordNet, these wordnets are interconnected with links stored in the Inter-Lingual Index (ILI). Following the EuroWordNet architecture, the MEANING project has developed the first versions of Multilingual Central Repository (MCR) using WordNet 1.6 as ILI. Thus, maintaining the compatibility between wordnets of different languages and versions. This version of the MCR integrates six different versions of the English WordNet (1.6 to 3.0) and wordnets in Spanish, Catalan, Basque and Italian, along with more than a million semantic relationships between concepts and semantic properties different ontologies. We recently developed a new version of MCR using WordNet 3.0 as ILI. This new version of the MCR integrates wordnets of five different languages: English, Spanish, Catalan, Basque and Galician. The current version of MCR, like the previous one, systematically integrates thousands of semantic relations between concepts. In addition, the MCR is enriched with about 460,000 semantic and ontological properties including *Base Level Concepts*, *Top Ontology*, *WordNet Domains* and *AdimenSUMO*, providing all ontological consistency the integrated semantic wordnets and resources on it.

**Keywords**

Lexical Semantics, Lexical Knowledge Bases, WordNet, EuroWordNet

## 1 Introducción

A pesar del progreso realizado en los últimos años en el área del Procesamiento del Lenguaje Natural (PLN), aún estamos lejos de comprender automáticamente textos en lenguaje natural. El uso de bases de conocimiento de amplia cobertura es una práctica común en los sistemas de PLN avanzados. Sin duda, la base de conociminto más utilizada es WordNet[1] (Fellbaum, 1998). No obstante, la construcción de bases de conocimiento con cobertura suficiente pare procesar textos de dominio general requiere de un esfuerzo enorme. Este esfuerzo sólo pueden realizarlo grandes grupos de investigación durante largos periodos de desarrollo. Por ejemplo, en el caso del WordNet desarrollado en Princeton para el inglés, en más de diez años de construcción manual (desde 1995 hasta 2006, es decir, de la versión 1.5 a la 3.0) creció de 103.445 a 235.402 relaciones semánticas[2], lo que representa un crecimiento de aproximadamente mil nuevas relaciones por mes. Sin embargo, en 2008, el grupo de Princeton distribuyó un nuevo recurso con 458.825 palabras de las definiciones de WordNet, manualmente anotadas con el correspondiente sentido de WordNet[3]. Afortunadamente, en los ultimos años, la comunidad investigadora ha desarrollado un amplio conjunto de recursos semánticos de amplia cobertura vinculados a distantas versiones de WordNet. A lo largo de los últimos años, muchos de estos recursos han sido integrados en el *Multilingual Central Repository* (MCR) (Atserias et al., 2004; Gonzalez-Agirre, Laparra e Rigau, 2012a; Gonzalez-Agirre, Laparra e Rigau, 2012b). El MCR sigue el modelo propuesto por el proyecto europeo EuroWordNet[4] (LE-2 4003) (Vossen, 1998). EuroWorNet diseñó una base de la datos lexical multilingüe con wordnets de varios idiomas europeos, estructuras de forma analoga al WordNet inglés. La versión actual del MCR es el resultado del Proyecto Europeo MEANING[5] (IST-2001-34460) (Rigau et al., 2002), así como de los proyectos KNOW[6] (TIN2006-15049-C03), KNOW2[7] (TIN2009-14715-C04) y de varias acciones complementarias asociadas al proyecto KNOW2.

El artículo está estructurado como sigue: En la sección 2 realizamos un repaso de las bases de conocimiento léxico existentes, introduciendo también la primera versión del *Multilingual Central Repository* (MCR). A continuación la sección 3 presenta la última versión del MCR y el *Web EuroWordNet Interface* (WEI), incluyendo una detallada descripción de la estructura de la base de satos empleada para implementar el MCR. Por último, en la sección 4 redactamos algunas conclusiones, y marcamos el camino para trabajos futuros.

## 2 Bases de Conocimiento Léxicas

Este sección proporciona una revisión de las bases de conocimiento léxico para el Procesamiento de Lenguaje Natural (PLN). La sección esta dividida en tres partes. Primero, el apartado 2.1 revisa los conceptos más importantes relacionados con las tareas de PLN, y el uso de los *recursos semánticos* de amplia cobertura. El siguiente apartado presenta las principales metodologías, estrategias y técnicas empleadas para la *construcción manual de recursos semánticos de gran tamaño* (apartado 2.2). Finalmente, el apartado 2.3 presenta la primera versión del Multilingual Central Repository (MCR).

### 2.1 Conocimiento Léxico y PLN

En el contexto del Procesamiento del Lenguaje Natural (PLN), la semántica estudia el significado, y en concreto se centra en la relación entre significantes, tales como las palabras, las frases, los signos y símbolos. En particular, la *Semántica Léxica* estudia el significado individual de las palabras y sus relaciones. La semántica léxica también estudia como está organizado el léxico y como el significado léxico esta interrelacionado. Su mayor objetivo es estructurar un modelo de léxico a través de la categorización de tipos de relación entre palabras. La semántica léxica se centra en el estudio de las unidades léxicas. Las unidades léxicas son los elementos básicos de un léxico (el vocabulario) y pueden ser consideradas como la unidad mínima de significado.

### 2.2 Construcción manual de Bases de Conocimiento

La tarea de Procesamiento de Lenguaje Natural (PLN) requiere de enormes bases de conocimiento semántico como respaldo de procesos semánticos intensos. Por ello, en los últimos años el desarrollo de recursos léxicos y semánticos de amplia cobertura ha sido un objetivo prioritario de investigación.

[1] `http://wordnet.princeton.edu`
[2] Las relaciones simétricas sólo se contabilizan una vez.
[3] `http://wordnet.princeton.edu/glosstag.shtml`
[4] `http://www.illc.uva.nl/EuroWordNet`
[5] `http://nlp.lsi.upc.edu/projectes/meaning`
[6] `http://ixa.si.ehu.es/know`
[7] `http://ixa.si.ehu.es/know2`

La construcción de estas bases de conocimiento requiere del esfuerzo de grandes grupos de investigación a lo largo de periodos de desarrollo prolongados. Sin embargo, estas bases de conocimiento, aún hoy en día, no parecen ser lo suficientemente ricos como para se empleados directamente en aplicaciones semánticas avanzadas. Parece que las aplicaciones de PLN no podrán mejorar sin la incorporación de conocimiento más detallado, rico y de propósito general.

Es más, todos los idiomas encapsulan el conocimiento de modos distintos. Esta variación entre idiomas es uno de los problemas principales que impide el uso extendido de las tecnologías de PLN. Una de las soluciones propuestas es la de adoptar una representación conceptual común que factorice la variación dentro de un idioma y también con el resto de los idiomas.

La necesidad de grandes bases de conocimiento semántico se puede vislumbrar observando la cantidad de proyectos que actualmente construyen recursos de este tipo. Proyectos como WordNet[8] (Fellbaum, 1998), FrameNet[9] (Baker, Fillmore e Lowe, 1998), VerbNet[10] (Kipper et al., 2006), SUMO[11] (Niles e Pease, 2001) o Cyc[12] (Lenat, 1995) han dedicado décadas y miles de horas de trabajo a la construcción manual de estos recursos de conocimiento semántico. Por desgracia, la construcción manual de estos recursos limita de forma severa su cobertura y escala.

Por ejemplo, la gran mayoría de ontologías formales se han desarrollado para dominios particulares[13]. Las ontologías son representaciones formales de un conjunto de conceptos dentro de un dominio, y de la relaciones entre dichos conceptos, normalmente incluyendo una taxonomía y un conjunto de relaciones semánticas. Las ontologías suelen ser empleadas para razonar sobre las propiedades de dichos dominios, y también pueden ser usadas para definir el dominio en sí (Álvez, Lucio e Rigau, 2012).

### *2.2.1 WordNet*

**WordNet**[14] (Miller et al., 1991; Fellbaum, 1998) es una base de conocimiento léxica para el idioma inglés. Esta inspirada por teorías psicolingüisticas y computacionales sobre la memoria léxica humana. Contiene información codificada manualmente sobre nombres, verbos, adjetivos y adverbios del inglés, y esta organizada entorno a la noción de *synset*. Un synset es un conjunto de palabras de la misma categoría morfosintáctica que pueden ser intercambiados en un contexto dado. Por ejemplo, $\langle student, pupil, educatee \rangle$ forman un synset porque pueden ser utilizados para referirse al mismo concepto. Un synset es comúnmente descrito por una *gloss* o definición, que en el caso del synset anterior es "*a learner who is enrolled in an educational institution*", y además, por un conjunto explícito de relaciones semánticas con otros synsets. Cada synset representa un concepto que está relacionado con otros conceptos mediante una gran variedad de relaciones semánticas, incluyendo hiperonimia/hiponimia, meronimia/holonimia, antonimia, etc. Los synsets están enlazados entre ellos mediante relaciones léxicas y semántico-conceptuales. WordNet también codifica 26 tipos diferentes de relaciones semánticas. WordNet está disponible de modo público y gratuito para su descarga. La versión actual de WordNet es la 3.1. Su estructura lo convierte en una herramienta útil para la lingüística computacional y el procesamiento de lenguaje natural. Resulta evidente que WordNet se ha convertido en un estándar en el PLN. De hecho, WordNet es usado en todo el mundo como base para anclar distintos tipos de conocimiento semántico, incluyendo wordnets de otros idiomas (Vossen, 1998), conocimiento de dominios (Magnini e Cavaglià, 2000) u ontologías como la Top Ontology (Álvez et al., 2008) o AdimenSUMO (Álvez, Lucio e Rigau, 2012).

WordNet ha sido creado y está siendo mantenido por el *Cognitive Science Laboratory* de la Universidad de Princeton inicialmente bajo la dirección del profesor George A. Miller y actualmente por la profesora Christiane D. Fellbaum. Su desarrollo comenzó en 1985. A lo largo de los años el proyecto ha recibido financiación de diferentes agencias del gobierno americano. WordNet ha sido empleado en una amplia variedad de tareas de PLN, tales como *Information Extraction* (Stevenson e Greenwood, 2006), *Automatic Summaritzation* (Chaves, 2001), *Question Answering* (Moldovan e Rus, 2001), *Lexical Expansion* (Parapar, Barreiro e Losada, 2005), etc.

La difusión y el éxito de WordNet ha provocado la aparición de multitud de proyectos con el objetivo de construir wordnets para otros idiomas, tomando como referencia la versión inglesa. Por ejemplo, catalán (Benítez et al., 1998), castellano (Atserias et al., 1997), euskera (Agirre et

[8] http://wordnet.princeton.edu

[9] http://framenet.icsi.berkeley.edu

[10] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

[11] http://www.ontologyportal.org/

[12] http://www.cyc.com

[13] http://protegewiki.stanford.edu/wiki/Protege_Ontology_Library

[14] http://wordnet.princeton.edu/

al., 2002), árabe (Rodríguez et al., 2008), etc. [15]

Algunos esfuerzos se han centrado en el desarrollo de wordnets multilingües, como EuroWordNet[16] (Vossen, 1998), MultiWordNet[17] (Pianta, Bentivogli e Girardi, 2002), Balkanet (Stamou et al., 2002b) o más recientemente, el WordNet Asiático[18] (Sornlertlamvanich, Charoenporn e Isahara, 2010), o wordnets de dominios particulares, como EuroTerm (Stamou et al., 2002a) o JurWordNet[19] (Sagri, Tiscornia e Bertagna, 2004).

La *Global WordNet Association*[20] es una organización sin fines lucrativos que proporciona un marco para establecer contactos, compartir y discutir sobre los wordnets que se desarrollan en todos los idiomas del mundo.

### *2.2.2 EuroWordNet*

El proyecto EuroWordNet[21] (Vossen, 1998) diseñó una arquitectura completa para el desarrollo de una base de conocimiento multilingüe que incluyera varios wordnets de idiomas europeos (entre ellos, holandés, italiano, castellano, alemán, francés, checo y estonio). En EuroWordNet, cada WordNet representa un único sistema interno de lexicalizaciones siguiendo la estructura del wordnet inglés. Los wordnets de los distintos idiomas están ligados mediante el *Inter-Lingual Index* (abreviado como ILI). Estas conexiones permiten acceder a palabras similares en cualquiera de los idiomas integrados en el arquitectura EuroWordNet. Además, el ILI dá acceso a una ontología lingüística compuesta por 63 relaciones semánticas distintas. Esta ontología proporciona una categorización común para todos los idiomas, mientras que las distinciones específicas de cada idioma están en cada uno de los wordnets locales.

Aunque el proyecto EuroWordNet se concluyó en el verano de 1999, muchos de sus principios siguen aún vigentes. Por ejemplo, el diseño de la arquitectura multilingüe, los *Base Concepts*, las relaciones, la ontología, etc. se han seguido usando por grupos de investigación que están desarrollando wordnets en otros idiomas (como por ejemplo, el castellano, el euskera, el catalán y el gallego) usando buena parte de la especificación de EuroWordNet. Si los wordnets son compatibles con la especificación, pueden ser añadidos a una base de datos común, y mediante el ILI, ser conectados con otros wordnets, permitiendo el uso aplicaciones multilingües de lenguaje natural.

### *2.2.3 Base Concepts*

The noción de los *Base Concepts* (a partir de ahora BC) fue introducida en EuroWordNet. Se supone que los BC son conceptos que juegan un papel importante en los diversos wordnets de diferentes idiomas. Este rol puede ser definido mediante dos criterios principales:

- Una posición elevada en la jerarquía semántica.
- Tener muchas relaciones con otros conceptos.

Por lo tanto, los BC son los bloques fundamentales para el establecimiento de relaciones en un wordnet y dar información acerca de los patrones dominantes de lexicalización en los idiomas. De este modo, los *Lexicografic Files* (o Supersentidos) de WordNet pueden ser considerados como el conjunto más básico de BC. Siguiendo estos criterios, en EuroWordNet se seleccionó un conjunto de BC para que se alcanzara un máximo de cobertura y compatibilidad durante el desarrollo de los wordnets de los distintos idiomas. Inicialmente, se seleccionó un conjunto de 1.024 *Common Base Concepts* extraídos de WordNet 1.5 (conceptos que actúan como BC en al menos dos idiomas), considerando solamente los wordnets en inglés, holandés, español e italiano.

Los *Basic Level Concepts* (Rosch e Lloyd, 1978) (a partir de ahora BLC) son el resultado de un compromiso entre dos principios de caracterización opuestos:

- Representar tantos conceptos como sea posible.
- Representar tantas características como sea posible.

Así, los BLC típicamente deberían ocurrir en niveles de abstracción medios, es decir, en posiciones intermedias de las jerarquías. Con esta idea en mente, diseñamos un algoritmo que utiliza propiedades estructurales básicas de cualquier versión de WordNet para obtener un conjunto completo de BLC que represente a todos sus sustantivos y verbos (Izquierdo, Suárez e Rigau, 2007)[22]. Para seleccionar los BLCs de forma

[15] Una lista de wordnets actualmente en desarrollo puede encontrarse en `http://www.globalwordnet.org/gwa/wordnet_table.html`

[16] `http://www.illc.uva.nl/EuroWordNet/`

[17] `http://multiwordnet.fbk.eu/`

[18] `http://www.asianwordnet.org`

[19] `http://www.ittig.cnr.it/Ricerca/materiali/JurWordNet/JurWordNetEng.htm`

[20] `http://www.globalwordnet.org`

[21] `http://www.illc.uva.nl/EuroWordNet/`

[22] `http://adimen.si.ehu.es/web/BLC`

automática, el programa calcula el número total de relaciones del *synset* o el número de relaciones de hiponímia y descarta los BLCs que no representan al menos un número determinado de *synsets* descendientes. Estos BLCs automáticos se han utilizado para *Word Sense Disambiguation* basada en clases semánticas (Izquierdo, Suárez e Rigau, 2009; Izquierdo, Suárez e Rigau, 2010) y para facilitar la conexión de WordNet con la ontología del proyecto KYOTO (Laparra, Rigau e Vossen, 2012).

### *2.2.4 Top Ontology*

Para maximizar un desarrollo uniforme y consistente de los wordnets, el proyecto EuroWordNet categorizó los *Base Concepts* usando la *Top Ontology*, que fue diseñado específicamente para este propósito. La **Top Ontology**[23] (Rodríguez et al., 1998) esta basada en clasificaciones lingüísticas ya existentes y adaptada para representar la diversidad de los *Base Concepts*. Es importante tener en cuenta que los *Top Concepts* representan características semánticas que puede ser aplicadas de forma conjuntiva. Por ejemplo, es posible obtener grupos complejos de caracteristicas, como *Container+Living+Part+Solid*, que puede ser aplicado, por ejemplo, a un "vaso sanguíneo".

El primer nivel de la *Top Ontology* está dividido en tres tipos:

- 1stOrderEntity (corresponde a objetos y sustancias concretas y perceptibles)
- 2ndOrderEntity (estados, situaciones y eventos)
- 3rdOrderEntitiy (entidades mentales como las ideas, conceptos y conocimientos)

Así, la *Top Ontology* de EuroWordNet está organizada mediante 63 características que pueden ser combinadas. La ontología esta especialmente diseñada para ayudar en la codificación de las relaciones léxico-semánticas en WordNet. Sin embargo, durante el proyecto EuroWordNet sólo se pudieron caracterizar con etiquetas de la TO los Base Concepts (BC) seleccionados en el proyecto.

Muchas de las subdivisiones de la TO son disjuntas. Por ejemplo, un concepto no puede ser a la vez *Natural* y *Artifact*. Explotando estas incompatibilidades entre las características de la TO podemos localizar inconsistencias ontológicas en la jerarquía de WordNet. Para ello, simplemente debemos heredar las características de la TO asignadas a los BC a través de la jerarquía de hiponimia de WordNet. Para evitar la herencia de categorías incompatibles podemos incluir algunos puntos de bloqueo en la jerarquía de hiponimia. De esta forma, hemos desarrollado un conjunto de herramientas para el control de la consistencia de la anotación y obtener su expansión. Para demostrar la consistencia de la anotación, hemos comprobado que no hay incompatibilidad en la anotación de la parte nominal de WordNet 1.6 cuando se utilizan los puntos de bloqueo. La expansión de la anotación se puede obtener cuando la anotación es consistente. Siguiendo este proceso hemos obtenido una anotación consistente de la parte nominal de WorNet[24] (Álvez et al., 2008).

### *2.2.5 WordNet Domains*

Uno de los problemas de WordNet es su nivel de granularidad. Hay conceptos cuyas diferencias a nivel semántico son virtualmente indetectables. **WordNet Domains**[25] (WND) (Magnini et al., 2002) es un recurso léxico desarrollado en el ITC-IRST por (Magnini e Cavaglià, 2000) donde los synsets han sido anotados de un modo semi-automático con una o más etiquetas de dominio, escogidas de un conjunto de 165 etiquetas organizadas jerárquicamente. Los usos de WND incluyen el poder de reducir el nivel de polisemia de las palabras y agrupar aquellos sentidos que pertenecen al mismo dominio. Por ejemplo, para la palabra *bank* (banco, en inglés), siete de sus diez sentidos en WordNet no comparten dominio, reduciendo de este modo la polisemia. Además, un dominio puede incluir *synsets* de diferentes categorías morfosintácticas. Por ejemplo, *MEDICINE* puede contener sentidos de nombres y de verbos. Un dominio también puede incluir sentidos de diferentes sub-jerarquias de WordNet. Por ejemplo *SPORTS* tiene conceptos subclase de *lifeform*, *physical-object*, *act*, *location*, etc. Sin embargo, la construcción de WND ha seguido un proceso semi-automático y aunque su anotación ha sido revisada (Bentivogli et al., 2004), aún podemos encontrar fácilmente muchas inconsistencias en su anotación (Castillo, Real e Rigau, 2004). Es por ello que se han propuesto y desarrollado métodos más robustos para la asignación de etiquetas de dominio a través de WordNet (González, Rigau e Castillo, 2012; Gonzalez-Agirre, Castillo e Rigau, 2012).

[23] `http://www.illc.uva.nl/EuroWordNet/corebcs/ewnTopOntology.html`

[24] `http://adimen.si.ehu.es/web/WordNet2TO`

[25] `http://wndomains.fbk.eu/`

### *2.2.6 SUMO y AdimenSUMO*

**SUMO**[26] (Niles e Pease, 2001) fue creado por el *IEEE Standard Upper Ontology Working Group*. Su objetivo era desarrollar una ontología estándard de alto nivel para promover el intercambio de datos, la búsqueda y extracción de información, la inferencia automática y el procesamiento del lenguaje natural. SUMO provee definiciones para términos de propósito general resultantes de fusionar diferentes ontologías libres de alto nivel (ej. la ontología de alto nivel de Sowa, axiomas temporales de Allen, mereotopología formal de Guarino, etc.).

SUMO consiste en un conjunto de conceptos, relaciones y axiomas que formalizan una ontología de alto nivel. Una ontología de alto nivel está limitada a conceptos que son meta, genéricos o abstractos. Por tanto, estos conceptos son suficientemente genéricos como para caracterizar un amplio rango de dominios. Aquellos conceptos que son de dominios especificos o particulares no están incluidos en una ontología de alto nivel.

SUMO está organizada en tres niveles. La parte superior y la parte central consisten en aproximadamente 1.000 términos y 4.000 axiomas, dependiendo de la versión. El tercer nivel contiene ontologías de dominio. En total, cuando todas las ontologías de dominio son combinadas, SUMO consiste en aproximadamente 20.000 términos y cerca de 70.000 axiomas.

Además, los desarrolladores de SUMO han creado un enlace completo a WordNet (Niles e Pease, 2003).

AdimenSUMO[27] (Álvez, Lucio e Rigau, 2012) es una reconversión de SUMO a una ontología de primera orden operativa. Así, AdimenSUMO puede ser utilizada para el razonamiento formal por demostradores de teoremas de lógicas de primer orden (como E-prover o Vampire). Al estar también enlazado a WordNet, AdimenSUMO se convierte en una herramienta muy potente para realizar razonamiento avanzado. Por ejemplo, utilizando demostradores de teoremas avanzados, es fácil inferir de AdimenSUMO que ninguna planta tiene cerebro (ni otras partes de animal).

## 2.3 Multilingual Central Repository

Uno de los principales resultados del proyecto MEANING[28] fue el desarrollo de la primera versión del Multilingual Central Repository (MCR)[29] (Atserias et al., 2004) para mantener la compatibilidad entre wordnets de distintos idiomas y versiones, tanto nuevos como anteriores, así como el nuevo conocimiento que se fuera adquiriendo.

Todo el diseño del MCR sigue la arquitectura propuesta por EuroWordNet. Esta arquitectura hace posible desarrollar wordnets locales de forma relativamente independiente, garantizando al mismo tiempo un alto nivel de compatibilidad. Esta estructura multilingüe permite transportar el conocimiento de un wordnet al resto de wordnets a través del ILI (*Inter-Lingual Index*), manteniendo la compatibilidad entre todos ellos. De esta forma, la estructura del ILI (incluyendo la *Top Ontology* (Vossen et al., 1997), *Wordnet Domains* (Magnini e Cavaglià, 2000) y la ontología *SUMO* (Niles e Pease, 2001)) actúa como la columna vertebral que permite transferir el conocimiento adquirido de cada uno de los wordnets locales al resto. Del mismo modo, los diferentes recursos (ej. las diferentes ontologías) están relacionadas mediante el ILI, y en consecuencia también pueden ser validados entre ellos (Álvez et al., 2008; Álvez, Lucio e Rigau, 2012).

El MCR sólo incluye conocimiento conceptual. Esto significa que tan solo las relaciones semánticas entre *synsets* pueden ser integradas y transportadas entre los diferentes wordnets. Aún así, cuando sea necesario, las relaciones adquiridas pueden mantenerse sub-especificadas. En ese sentido, pueden integrarse y transportarse a otros idiomas o procesos. Por ejemplo, la relación *<gain> involved <money>* capturada como un objeto-directo típico, más tarde puede detallarse como *<gain> involved-patient <money>* y ser portada al wordnet en castellano como *<ganar> involved-patient <dinero>*.

La versión del MCR desarrollada en el marco del proyecto MEANING contiene seis versiones distintas del WordNet inglés (Fellbaum, 1998) (de la 1.5 a la 3.0) junto con más de un millon de relaciones semánticas entre *synsets* adquiridas de WordNet, eXtended WordNet (Mihalcea e Moldovan, 2001), y preferencias de selección adquiridas de *SemCor* (Agirre e Martínez, 2001; Agirre e Martínez, 2002) y del *British National Corpus* (BNC) (McCarthy, 2001). Esta versión del MCR también incluye wordnets del castellano (Atserias et al., 1997), italiano (Bentivogli, Pianta e Girardi, 2002), euskera (Agirre et al., 2002) y catalán (Benítez et al., 1998). Esta versión usa un ILI basado en WordNet 1.6.

Como estos recursos han sido desarrollados usando diferentes versiones de WordNet (de la

[26] `http://www.ontologyportal.org`
[27] `http://adimen.si.ehu.es/web/AdimenSUMO`
[28] `http://nlp.lsi.upc.edu/projectes/meaning`
[29] `http://adimen.si.ehu.es/web/MCR`

1.5 a la 3.0), hemos tenido que aplicar una tecnología que alineara los wordnets automáticamente **WordNet Mappings**[30] (Daudé, 2005). Esta tecnología proporciona enlaces entre synsets de diferentes versiones de WordNets, manteniendo la compatibilidad de todos los recursos que usan una determinada versión de WordNet. Además, esta tecnología permite realizar el transporte de todo el conocimiento asociado a una version de WordNet al resto de versiones.

Al término del proyecto MEANING, el MCR ha continuado su desarrollo y mejora en los proyectos nacionales KNOW[31] y KNOW2[32], así como varias acciones complementarias, con especial énfasis en los idiomas inglés, castellano, catalán, euskera y gallego.

La versión actual del MCR integra, siguiendo la arquitectura EuroWordNet, wordnets de cinco idiomas diferentes: inglés, castellano, catalán, euskera y gallego. El *Inter-Lingual-Index* (ILI) permite la conectividad entre las palabras en un idiomas con las traducciones equivalentes en cualquiera de las otras lenguas gracias a los enlaces generados automáticamente. El ILI actual corresponde a la versión 3.0 de WordNet.

Por ello, el MCR constituye un recurso multilingüe de amplia cobertura que puede ser de gran utilidad para un gran número de procesos semánticos que requieren de conocimientos lingüístico-semánticos ricos y complejos (por ejemplo, ontologías para la web semántica). Así, el MCR está siendo utilizado en múltiples proyectos y desarrollos. Por ejemplo, los proyectos europeos KYOTO[33], PATHS[34], OpeNER[35] y NewsReader[36], y el proyecto nacional SKaTer[37].

### *2.3.1 MCR usando ILI 1.6*

La versión del MCR que usa un ILI basado en WordNet 1.6 tiene los siguientes componentes:

- ILI (versión WordNet 1.6):
  - WordNet 1.6 (Fellbaum, 1998)
  - Base Concepts (Izquierdo, Suárez e Rigau, 2007)
  - Top Ontology (Álvez et al., 2008)
  - WordNet Domains (Bentivogli et al., 2004)
  - AdimenSUMO (Álvez, Lucio e Rigau, 2012)
- Wordnets locales:
  - WordNet inglés: versiones 1.5, 1.6, 1.7.1, 2.0, 2.1, 3.0 (Fellbaum, 1998)
  - wordnets castellano y catalán (Benítez et al., 1998), italiano (Bentivogli, Pianta e Girardi, 2002) y euskera (Agirre et al., 2002).
  - eXtended WordNet (Mihalcea e Moldovan, 2001)
- Preferencias semánticas:
  - Adquiridas de SemCor (Agirre e Martínez, 2002)
  - Adquiridas del BNC (McCarthy, 2001)
- Instancias
  - Entidades nombradas (Alfonseca e Manandhar, 2002)

Inicialmente, la mayor parte del conocimiento que pretendíamos integrar en el MCR estaba alineado a WordNet 1.6, el WordNet italiano o de *MultiWordNet Domains*, éstos últimos desarrollados usando un ILI basado en WordNet 1.6 (Bentivogli, Pianta e Girardi, 2002; Magnini e Cavaglià, 2000). Por tanto, el MCR usando un ILI basado en WordNet 1.6 minimizaba efectos secundarios con otras iniciativas europeas (proyectos Balkanet, EuroTerm, etc.) y otros wordnets desarrollados alrededor de la *Global WordNet Association*. Sin embargo, el ILI para los wordnets del castellano, catalán y euskera era el WordNet 1.5 (Atserias et al., 1997; Benítez et al., 1998), así como la *Top Ontology* de EuroWordNet y los *Base Concepts* asociados. Por tanto, éstos últimos recursos debieron transportarse a la versión WordNet 1.6 (Atserias, Villarejo e Rigau, 2003). Además, la versión final del MCR con ILI basado en WordNet 1.6 terminada al final del proyecto KNOW2 contiene versiones mejoradas de *Base Concepts* (Izquierdo, Suárez e Rigau, 2007), *Top Ontology* (Álvez et al., 2008), *WordNet Domains* (Bentivogli et al., 2004) y *AdimenSUMO* (Álvez, Lucio e Rigau, 2012). Sin embargo, muchos de sus componentes tenían licencias restrictivas y no podían distribuirse de forma libre e integrada con el resto del MCR[38]. Por ello, y para actualizar los recursos existentes decidimos actualizar el ILI a WordNet 3.0.

[30] `http://www.talp.upc.edu/index.php/technology/resources/multilingual-lexicons-and-machine-translation-resources/multilingual-lexicons/98-wordnet-mappings`

[31] `http://ixa2.si.ehu.es/know`

[32] `http://ixa2.si.ehu.es/know2`

[33] `http://www.kyoto-project.eu`

[34] `http://www.paths-project.eu`

[35] `http://www.opener-project.org/`

[36] `http://www.newsreader-project.eu/`

[37] `http://nlp.lsi.upc.edu/skater`

[38] Esta versión puede consultarse en `http://adimen.si.ehu.es/cgi-bin/wei1.6/public/wei.consult.perl`

## 3 Multilingual Central Repository 3.0

La versión actual del MCR usa un ILI basado en WordNet 3.0 e integra, siguiendo el modelo propuesto por EuroWordNet y MEANING, wordnets de cinco idiomas distintos, incluyendo el inglés, castellano, catalán, euskera y gallego. Como en la versión anterior, los wordnets están conectados a través del *Inter-Lingual-Index* (ILI) permitiendo conectar palabras de un idioma con las palabras equivalentes en los otros idiomas también integrados en el MCR. Como la versión actual del ILI del MCR es la correspondiente a la versión 3.0 del WordNet en inglés, la mayoría del conocimiento ontológico ha sido transportado desde las versiones anteriores al nuevo MCR 3.0. Así, la mayoría de los recursos transportados han tenido que ser alineados a la nueva versión. La descripción completa del proceso empleado para llevar a cabo el transporte y actualización de todos los recursos se puede consultar en (Gonzalez-Agirre, Laparra e Rigau, 2012b; Gonzalez-Agirre, Laparra e Rigau, 2012a).

Además, para poder interactuar con el MCR y actualizar su contenido, hemos actualizado el *Web EuroWordNet Interface* (WEI), una nueva interfaz web para navegar y editar el MCR 3.0.

### 3.1 Web EuroWordNet Interface

El *Web EuroWordNet Interface* (WEI) (Benítez et al., 1998) permite consultar y editar la información contenida en el MCR. WEI usa technología CGI, lo que significa que todos los datos se procesan sólo en el servidor y los usuarios trabajan con clientes ligeros con capacidad de navegación web HTML. Todos los datos se almacenan en una base de datos MySQL. La interfaz se ha ido actualizando desde su desarrollo inicial en el proyecto EuroWordNet.

WEI permite navegar, consultar y editar la información asociada a un item (que puede ser un *synset*, una palabra, un *variant* o un ILI) de uno de los wordnets integrados en el MCR. La aplicación WEI consulta la información correspondiente a ese item y la información ontológica asociada a los ILIs correspondientes. La aplicación también permite consultar por los *synsets* relacionados del propio wordnet origen de la consulta (usando las relaciones codificadas en el MCR de hiperonimia, hiponimia, meronimia, etc.), o de algún otro wordnet integrado en el MCR (a través del ILI).

La aplicación consta de dos marcos. En el marco superior introducimos los parámetros para la búsqueda, y en inferior nos muestra los resultados de la consulta. Los diferentes parámetros de búsqueda son los siguientes:

- **Item:** el item que pretendemos buscar, que puede ser una palabra, un *synset*, un *variant* o un ILI.
- **Tipo de item:** el tipo del item que pretendemos buscar (palabra, *variant*, *synset* o ILI).
- **PoS:** la categoría gramatical del item (nombres, verbos, adjetivos o adverbios).
- **Relación:** que se carga dinámicamente desde la base de datos (sinónimos, hipónimos, hiperónimos, etc.)
- **WordNet origen:** el wordnet desde donde realizamos la consulta.
- **WordNet navegación:** el wordnet al cual seguimos las relaciones.
- **Glossa:** si está seleccionado se muestran las glosas de los *synsets*.
- **Score:** si está seleccionado se muestran los valores de confianza.
- **Rels:** si está seleccionado muestra información acerca de las relaciones que el *synset* tiene en todos los wordnets seleccionados.
- **Full:** si está seleccionado realiza una búsqueda transitiva por todas las relaciones.
- **WordNets mostrados:** wordnets seleccionados.

WEI también permite editar el contenido del MCR. Su funcionamiento es exactamente igual a la consulta, pero en modo edición, tanto los *synsets* como los ILIs pueden seleccionarse y editarse. Al editar un *synset* nos aparece una pantalla de donde podemos añadir, eliminar o modificar los *variants* del *synset*, modificar su glosa y ejemplos, así como las relaciones que tiene con otros *synsets* (en la Figura 1 se puede ver un ejemplo para el sentido 1 de "*party*"). Al editar un ILI nos aparece una pantalla donde podremos añadir, eliminar o modificar la información ontológica asociada al ILI.

#### *3.1.1 Marcas para synsets y variants*

En la nueva versión del WEI es posible asignar propiedades especiales o *marcas* a *variants* y a *synsets*. También podemos añadir una pequeña nota o comentario que especifica mejor por qué hemos asignado una marca determinada. Uno de los objetivos de las marcas es permitir una edición más rápida mediante WEI, siendo


Figura 1: Captura de pantalla de la interfaz de edición, mostrando el sentido 1 de "party".

capaces de marcar y anotar un synset, facilitando una posterior revisión y corrección. Otra ventaja es aumentar la cantidad de información almacenada en cada *variant* y *synset*.

Las marcas disponibles son las siguientes:

- Marcas de *variant*:
  - DUBLEX: Para aquellos *variants* con una lexicalización dudosa.
  - INFL: Indica que un *variant* es **flexivo**. Necesario para el wordnet en euskera.
  - RARE: *Variant* muy poco usado u en desuso.
  - SUBCAT: Subcategorización. Se usa para aquellos *variant* que deben tener subcategorización.
  - VULG: Para aquellos *variant* que son vulgares, rudos u ofensivos.
- Marcas de *synset*:
  - GENLEX: Conceptos generales no lexicalizados que son introducidos para organizar mejor la jerarquía.
  - HYPLEX: Indica que el hiperónimo tiene idéntica lexicalización.
  - SPECLEX: Termimos especificos de ciertos dominios, y que deben ser comprobados.

Anotar marcas usando el WEI es muy simple. Tan solo hay que buscar un synset o variant y anotarlo usando la interfaz de edición, seleccionando las marcas deseadas. En una sola ventana se pueden editar las marcas y comentarios de todos los variant contenidos en un synset (incluyendo más de uno, o incluso todos, a la vez), y la marca y el comentario del propio synset.

### 3.2 Diseño de la Base de Datos para el MCR

Actualmente, el MCR está almacenado en una base de datos relacional consistente de 40 tablas[39]. Este apartado describe cada una de las tablas necesarias para que el MCR funcione correctamente.

La tabla principal del MCR es la tabla que contiene el *Inter-Lingual Index* (ILI):

- **wei_ili_record**: Contiene los identificadores de los ILI, en formato ili-30-xxxxxxx-y (las $x$ indican el offset del *synset*, y las $y$ representan la categoria gramatical o *Part-Of-Speech* (PoS)). Cada registro también almacena el origen del ILI (el WordNet del que proviene), si es un *Base Concept* o no, el fichero lexicográfico, y si es una instancia o no.

Cada uno de los idiomas incluidos en el MCR (incluyendo el inglés) esta ligado al ILI, y compuesto por cinco tablas. Cada idioma tiene su propio código de 3 letras, que está indicado por *xxx* en la lista siguiente:

[39] La distribución actual ha sido probada sobre MySQL y PostgreSQL.

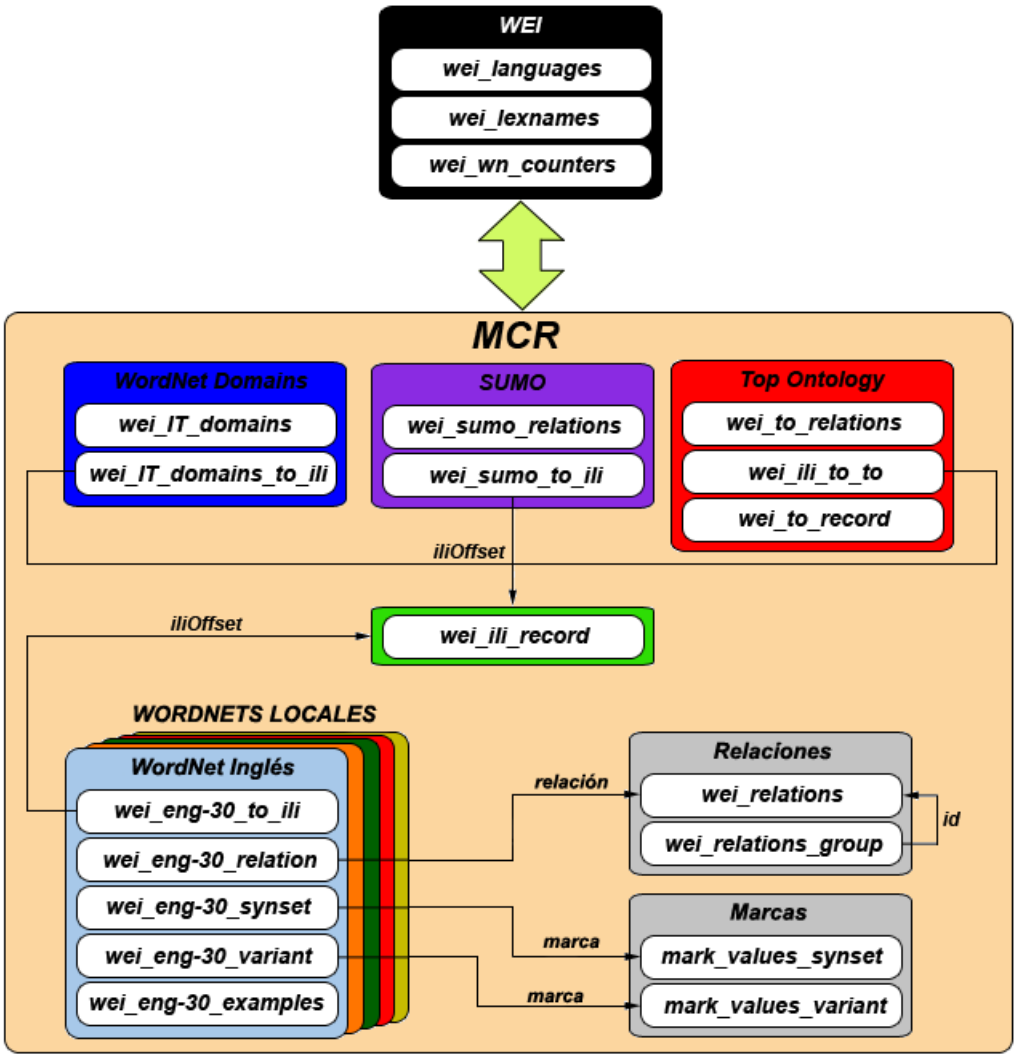


Figura 2: Estructura de la Base de Datos para el MCR y el WEI.

- **wei_xxx-30_to_ili**: Esta tabla conecta el ILI (por ejemplo, ili-30-00001740-a) con el número de *synset* (por ejemplo, eng-30-00001740-a).
- **wei_xxx-30_relation**: Esta tabla contiene todas las relaciones del wordnet. Cada registro (que es la instancia de una relación) guarda un código que indica el tipo de relación (que se almacena de la tabla *wei_relations*), la dirección de la relación (*synset* origen y *synset* destino), el valor de confianza y el wordnet del que proviene.
- **wei_xxx-30_synset**: Aquí se almacena la información acerca del *synset*: el identificador, el número de descendientes, la glosa, el nivel en el que se encuentra (contando desde arriba), y finalmente una marca (opcional) y un comentario del *synset* (opcional).
- **wei_xxx-30_variant**: Aquí se almacenan todos los *variant* del wordnet. Cada registro representa un único *variant* y contiene la siguiente información: el *variant*, el sentido de la palabra, el identificador de *synset*, un valor de confianza, el experimento del que proviene (opcional), y finalmente la marca (opcional) y el comentario del *variant* (opcional).
- **wei_xxx-30_examples**: En esta tabla se listan todos los ejemplos del wordnet. Cada ejemplo está identificado por el número de *synset*, la palabra y el sentido.

Añadir un nuevo idioma es tan fácil como crear las cinco nuevas tablas con el patrón anterior y un código de 3 letras que lo representa.

Además de los wordnets, el MCR integra otros recursos (dominios, ontologías, marcas, etc.). Las tablas que contienen esta información son las siguientes:

Dominios:

- **wei_domains**: Esta tabla representa la jerarquía de *WordNet Domains* usando tuplas origen-destino.
- **wei_ili_to_domains**: Cada registro enlaza un dominio a un ILI. También se indica el wordnet del que proviene. Esta tabla es única, haciendo que la información de dominios esté compartida entre todos los wordnets.

AdimenSUMO:

- **wei_sumo_relations**: Esta tabla representa la jerarquía de *AdimenSUMO* usando tuplas origen-destino. También incluye un campo que indica si se trata de una sub-clase o no.
- **wei_ili_to_sumo**: Cada registro enlaza una etiqueta de *AdimenSUMO* a un ILI. También se indica el wordnet del que proviene. Al igual que la tabla de dominios, esta tabla es única, haciendo que la información de dominios esté compartida entre todos los wordnets.

Top Ontology:

- **wei_to_relations**: Esta tabla representa la jerarquía de *Top Ontology* usando tuplas origen-destino. También incluye un campo que indica el tipo de la relación.
- **wei_ili_to_to**: Cada registro enlaza una etiqueta de *Top Ontology* a un ILI. También se indica el wordnet del que proviene. Al igual que la tabla de dominios y la de *AdimenSUMO*, esta tabla es única, haciendo que la información de dominios esté compartida entre todos los idiomas.
- **wei_to_record**: Esta tabla almacena, para cada etiqueta de *Top Ontology*, la glosa asociada a ella.

Marcas:

- **mark_values_synset**: Valores permitidos para las marcas de *synset*, asi como su descripción.
- **mark_values_variant**: Valores permitidos para las marcas de *variant*, asi como su descripción.

El resto de tablas incluidas en el MCR son las siguientes:

- **wei_relations**: Esta tabla contiene todas las relaciones posibles en el MCR. Cada relación tiene un identificador, un nombre, sus propiedades y una nota (opcional). También se indica si es inversa (en caso de que sea posible) y a que grupo de relaciones pertenece (ver más abajo). El código ID que aparece en esta tabla es el que está reflejado en las tabla *wei_xxx-30_relation*. Esta tabla es la que permite realizar búsquedas mediante el WEI.
- **wei_relations_group**: Aquí se almacenan los super-grupos de relaciones (sinónimos, hiperónimos, merónimos, causa, etc.). El código ID que aparece en esta tabla es el que está reflejado en la tabla *wei_relations*.
- **wei_languages**: Los wordnets disponibles en el MCR. Para cada wordnet se indica el código, el nombre y el color con el que debe de aparecer en el WEI.
- **wei_lexnames**: Aquí se almacenan los ficheros lexicográficos de WordNet. Cada entrada tiene un código (el indicado en la tabla *wei_ili_record*) y un nombre descriptivo.
- **wei_wn_counters**: La interfaz del WEI permite la creación de nuevos *synset*. Para evitar solapamientos y problemas futuros, cada PoS tiene su propio número de *offset*, empezando desde el 800.000. Esta tabla guarda los números que deben adoptar los nuevos *synsets* que se creen en cada categoría gramatical.

La figura 2 muestra la estructura completa del MCR.

### 3.3 Estado actual del MCR

En este apartado se presenta el estado actual del MCR, incluyendo el progreso respecto al WordNet inglés. La tabla 1 muestra la cantidad actual de *synsets* y *variants*, el número de glosas y el número de ejemplos de cada wordnet, distinguiendo entre las distintas categorías gramaticales o PoS.

Como ejemplo de la información contenida en el MCR podemos analizar el sentido 4 de "*party*", que se muestra en la figura 3. En la columna de la izquierda podemos ver el ILI (*ili-30-07447641-n*), y debajo de él, la información asociada al ILI:

- **WordNet Domains:** *free_time, sociology.*
- **Fichero semántico de WordNet:** *event.*
- **AdimenSUMO:** *Meeting.*
- **Top Ontology:** *Agentive, BoundedEvent, Communication, Purpose, Social.*

En la siguiente columna se muestra los *synsets* asociados de cada wordnet y los *variants* que hay

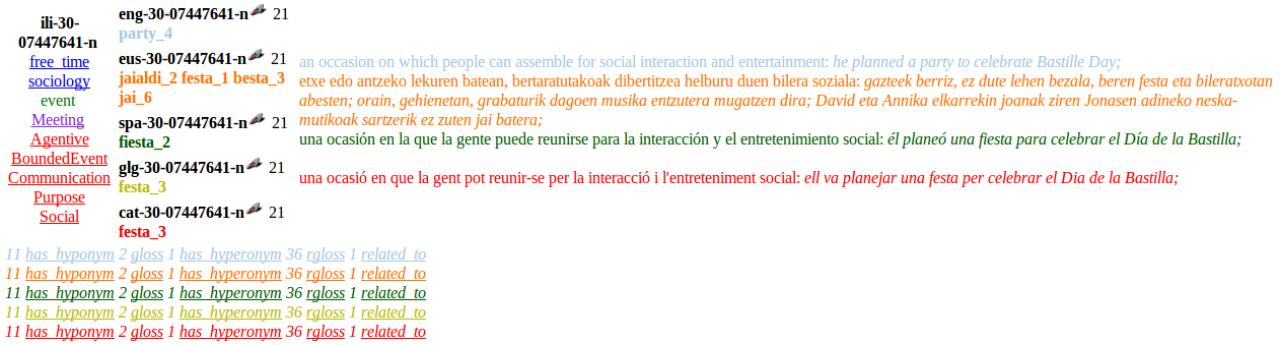


Figura 3: Información almacenada en el MCR para el sentido 4 de "party".

en cada uno de ellos, indicando el **variant** y el número de **sentido** de dicha palabra (el número final después del simbolo "_"):

- **Inglés:** party (4).
- **Euskera:** jaialdi (2), festa (1), besta (3), jai (6).
- **Castellano:** fiesta (2).
- **Gallego:** festa (3).
- **Catalán:** festa (3).

En la columna de la derecha están las glosas o definiciones para cada uno de los wordnets (en este caso, el gallego no tiene esta información), así como ejemplos de uso de las palabras (que se muestran en cursiva).

Por último, en la parte inferior se enumeran las relaciones de estos synset (que en este caso coinciden):

- **has_hyponym:** Indica que el synset tiene 11 hipónimos.
- **gloss:** Indica que 2 de las palabras que aparecen en la glossa estan desambiguadas, y ligadas al synset correspondiente. En este caso son *entertainment* e *interaction*, que podemos ver en la glossa en inglés.
- **has_hyperonym:** Indica que el synset tiene un hiperónimo.
- **rgloss:** *Reverse gloss*. Indica que este *synset* aparece desambiguado en 36 glosas de otros *synets*.
- **related_to:** Indica qué otros *synset* está relacionados con este. En este caso es el sentido 1 de "*party*" en inglés, definido como *have or participate in a party*, o el sentido 4 de "festejar" en castellano.

## 4 Conclusiones y Trabajo Futuro

Como hemos visto, la construcción de bases de conocimiento con cobertura suficiente pare procesar textos de dominio general requiere de un esfuerzo enorme. Este esfuerzo sólo pueden realizarlo grandes grupos de investigación durante largos periodos de desarrollo. Afortunadamente, en los ultimos años, la comunidad investigadora ha desarrollado un amplio conjunto de recursos semánticos de amplia cobertura vinculados a distantas versiones de WordNet. A lo largo de los últimos años, muchos de estos recursos han sido integrados en el *Multilingual Central Repository* (MCR). Primero usando un ILI basado en WordNet 1.6. Esta versión, aunque demostró el potencial de la propuesta contenía recursos con licencias restrictivas que impedían su distribución integrada. La versión que usa el ILI basado en WordNet 3.0 no contiene recursos que tengan licencias restrictivas y puede distribuirse de forma integrada[40].

Como vemos, integrar todos estos recursos en una única infraestructura también es una tarea compleja que requiere de un esfuerzo continuado. Por un lado, continuamente aparecen nuevos recursos potencialmente interesantes, y por otro, los antiguos recursos se siguen actualizando.

En el marco del proyecto SKaTer planeamos seguir enriqueciendo el MCR con nuevos recursos. Entre otros, los ya integrados en la versión del MCR con el ILI basado en WordNet 1.6. Por ejemplo, el resto de versiones de WordNet inglés (1.5, 1.6, 1.7, 1.7.1, 2.0 y 2.1, ya que resulta muy práctico poder consultar todas las versiones de WordNet simultáneamente). También pensamos recuperar y integrar al nuevo ILI las preferencias de selección adquiridas de SemCor, así como in-

[40] La mayor parte de los recursos integrados en esta versión tiene licencia Creative Commons Attribution 3.0 Unported (CC BY 3.0) `http://creativecommons.org/licenses/by/3.0`

| **Variants** | **Nombres** | **Verbos** | **Adjetivos** | **Adverbios** | **Synsets** | **%WN** |
|---|---|---|---|---|---|---|
| EngWN3.0 | 147.360 | 25.051 | 30.004 | 5.580 | 118.431 | 100 % |
| SpaWN3.0 | 39.142 | 10.824 | 6.967 | 1.051 | 38.702 | 33 % |
| CatWN3.0 | 51.605 | 11.577 | 7.679 | 2 | 46.033 | 39 % |
| EusWN3.0 | 40.939 | 9.470 | 148 | 0 | 30.615 | 26 % |
| GalWN3.0 | 18.949 | 1.416 | 6.773 | 0 | 19.312 | 16 % |
| **Glosas** | | | | | | |
| EngWN3.0 | 82.379 | 13.767 | 18.156 | 3.621 | 117.923 | 100 % |
| SpaWN3.0 | 12.533 | 3.325 | 1.917 | 670 | 18.445 | 16 % |
| CatWN3.0 | 6.294 | 44 | 840 | 1 | 7.179 | 6 % |
| EusWN3.0 | 2.690 | 2 | 0 | 0 | 2.692 | 2 % |
| GalWN3.0 | 4.997 | 2 | 3.111 | 0 | 8.111 | 7 % |
| **Ejemplos** | | | | | | |
| EngWN3.0 | 10.433 | 11.583 | 15.615 | 3.674 | 41.305 | 100 % |
| SpaWN3.0 | 465 | 30 | 195 | 193 | 606 | 2 % |
| CatWN3.0 | 2.105 | 46 | 368 | 0 | 2.201 | 5 % |
| EusWN3.0 | 2.376 | 0 | 0 | 0 | 2.075 | 5 % |
| GalWN3.0 | 270 | 2 | 4.291 | 0 | 2.416 | 6 % |

Cuadro 1: Número actual de *variants*, *synsets*, definiciones y ejemplos de cada wordnet.

formación sobre predicados verbales y nominales, tanto de VerbNet (Kipper et al., 2006), como de FrameNet (Baker, Fillmore e Lowe, 1998; Laparra, Rigau e Cuadros, 2010). Entre nuestras prioridades también está la integración y explotación de recursos para el análisis de sentimiento. Por ejemplo, Q-WordNet[41] (Agerri e García-Serrano, 2010).

## Agradecimentos

Este trabajo ha sido posible gracias al apoyo de los proyectos europeos MEANING (IST-2001-34460), KYOTO (ICT-2007-211423), OpeNER (ICT-2011-296451) y NewsReader (ICT-2011-316404), así como a los nacionales KNOW (TIN2006-15049-C03), KNOW2 (TIN2009-14715-C04-04), SKaTer (TIN2012-38584-C06-01), y varias acciones complementarias asociadas al proyecto KNOW2. Aitor GonzalezAgirre también recibe el apoyo del Ministerio Español de Educación, Cultura y Deporte a través de una beca pre-doctoral FPU (FPU12/06243).

## Bibliografía

Agerri, Rodrigo e Ana García-Serrano. 2010. Q-wordnet: Extracting polarity from wordnet senses. Em *Seventh Conference on International Language Resources and Evaluation, Malta (retrieved May 2010)*.

Agirre, Eneko, Olatz Ansa, Xabier Arregi, José Mª Arriola, Arantza Diaz de Ilarraza, E. Pociello, e L. Uria. 2002. Methodological issues in the building of the basque wordnet: quantitative and qualitative analysis. Em *Proceedings of the first International Conference of Global WordNet Association*, Mysore, India.

Agirre, Eneko e David Martínez. 2001. Learning class-to-class selectional preferences. Em *Proceedings of CoNLL01*, Toulouse, France.

Agirre, Eneko e David Martínez. 2002. Integrating selectional preferences in wordnet. Em *Proceedings of the 1st International Conference of Global WordNet Association*, Mysore, India.

Alfonseca, Enrique e S. Manandhar. 2002. Distinguishing concepts and instances in wordnet. Em *Proceedings of the first International Conference of Global WordNet Association*, Mysore, India, 21-25 January, 2002.

Atserias, Jordi, Salvador Climent, Javier Farreres, German Rigau, e Horacio Rodríguez. 1997. Combining multiple methods for the automatic construction of multilingual wordnets. Em *Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP'97)*, Tzigov Chark, Bulgaria.

Atserias, Jordi, Luís Villarejo, e German Rigau. 2003. Integrating and porting knowledge across languages. Em *Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP'03)*,

[41] http://www.rodrigoagerri.net/sentiment-analysis

pp. 31–37, Borovets, Bulgaria, September, 2003.

Atserias, Jordi, Luís Villarejo, German Rigau, Eneko Agirre, John Carroll, Piek Vossen, e Bernardo Magnini. 2004. The meaning multilingual central respository. Em *Proceedings of the Second International Global WordNet Conference (GWC'04)*.

Baker, Collin F., Charles J. Fillmore, e John B. Lowe. 1998. The berkeley framenet project. Em *Proceedings of the COLING-ACL*, pp. 86–90.

Bentivogli, Luisa, Pamela Forner, Bernardo Magnini, e Emanuele Pianta. 2004. Revising the wordnet domains hierarchy: semantics, coverage and balancing. Em *Proceedings of the Workshop on Multilingual Linguistic Ressources*, pp. 101–108. Association for Computational Linguistics.

Bentivogli, Luisa, E. Pianta, e C. Girardi. 2002. Multiwordnet: developing an aligned multilingual database. Em *Proceedings of the First International Conference on Global WordNet*, Mysore, India.

Benítez, Laura, Sergi Cervell, Gerard Escudero, Mónica López, German Rigau, e Mariona Taulé. 1998. Methods and tools for building the catalan wordnet. Em *Proceedings of ELRA Workshop on Language Resources for European Minority Languages*, Granada, Spain.

Castillo, Mauro, Francis Real, e German Rigau. 2004. Automatic assignment of domain labels to wordnet. Em *Proceeding of the 2nd International WordNet Conference*, pp. 75–82.

Chaves, R. P. 2001. Wordnet and automated text summarization. Em *Proceedings of 6th Natural Language Processing Pacific Rim Symposium NLPRS 2001*, pp. 109–116, Tokyo, Japan, Jan, 2001.

Daudé, Jordi. 2005. *Enlace de Jerarquías Usando el Etiquetado por Relajación*. Tese de doutoramento, Universitat Politècnica de Catalunya, July, 2005.

Fellbaum, Christiane. 1998. *WordNet. An Electronic Lexical Database*. Language, Speech, and Communication. The MIT Press.

González, Aitor, German Rigau, e Mauro Castillo. 2012. A graph-based method to improve wordnet domains. Em *Proceedings of the Computational Linguistics and Intelligent Text Processing (CICLING'12)*, pp. 17–28. Springer.

Gonzalez-Agirre, A., E. Laparra, e G. Rigau. 2012a. Multilingual central repository version 3.0. Em *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 2525–2529.

Gonzalez-Agirre, A., E. Laparra, e G. Rigau. 2012b. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. Em *Proceedings of the 6th Global WordNet Conference (GWC'12)*.

Gonzalez-Agirre, Aitor, Mauro Castillo, e German Rigau. 2012. A proposal for improving wordnet domains. Em Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, e Stelios Piperidis, editores, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may, 2012. European Language Resources Association (ELRA).

Izquierdo, R., A. Suárez, e G. Rigau. 2007. Exploring the automatic selection of basic level concepts. Em *Proceedings of RANLP*.

Izquierdo, Rubén, Armando Suárez, e German Rigau. 2009. An empirical study on class-based word sense disambiguation. Em *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 389–397. Association for Computational Linguistics.

Izquierdo, Rubén, Armando Suárez, e German Rigau. 2010. Gplsi-ixa: Using semantic classes to acquire monosemous training examples from domain texts. Em *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 402–406, Uppsala, Sweden, July, 2010. Association for Computational Linguistics.

Kipper, Karin, Anna Korhonen, Neville Ryant, e Martha Palmer. 2006. Extending verbnet with novel verb classes. Em *Proceedings of LREC*, volume 2006, pp. 1.

Laparra, Egoitz, German Rigau, e Montse Cuadros. 2010. Exploring the integration of wordnet and framenet. Em *Proceedings of the 5th Global WordNet Conference (GWC'10)*, Mumbai (India), January, 2010.

Laparra, Egoitz, German Rigau, e Piek Vossen. 2012. Mapping wordnet to the kyoto ontology. Em Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard,

Joseph Mariani, Jan Odijk, e Stelios Piperidis, editores, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may, 2012. European Language Resources Association (ELRA).

Lenat, Douglas B. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Magnini, Bernardo e G. Cavaglià. 2000. Integrating subject field codes into wordnet. Em *Proceedings of the Second Internatgional Conference on Language Resources and Evaluation (LREC)*, Athens. Greece.

Magnini, Bernardo, C. Satrapparava, G. Pezzulo, e A. Gliozzo. 2002. The role of domains informations. Em *In Word Sense Disambiguation*, Treto, Cambridge, July, 2002.

McCarthy, Diana. 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Aternations, Subcategorization Frames and Selectional Preferences*. Tese de doutoramento, University of Sussex.

Mihalcea, Rada e Dan Moldovan. 2001. extended wordnet: Progress report. Em *Proceedings of NAACL Workshop* WordNet and Other Lexical Resources: Applications, Extensions and Customizations, pp. 95–100, Pittsburg, PA, USA.

Miller, G. A., R. Beckwith, Christiane Fellbaum, D. Gross, K. Miller, e R. Tengi. 1991. Five papers on wordnet. *Special Issue of the International Journal of Lexicography*, 3(4):235–312.

Moldovan, Dan e Vasile Rus. 2001. Logic form transformation of wordnet and its applicability to question answering. Em *In Proceedings of ACL 2001*, pp. 394–401.

Niles, I. e Adam Pease. 2001. Towards a standard upper ontology. Em Chris Welty e Barry Smith, editores, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, Ogunquit, Maine.

Niles, I. e Adam Pease. 2003. Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. Em *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*, Las Vegas, Nevada.

Parapar, David, Alvaro Barreiro, e David E. Losada. 2005. Query expansion using wordnet with a logical model of information retrieval. Em *Proceedings of IADIS AC*, pp. 487–494.

Pianta, Emanuele, Luisa Bentivogli, e Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. Em *Proceedings of the First International Conference on Global WordNet*, January, 2002.

Rigau, German, Bernardo Magnini, Eneko Agirre, Piek Vossen, e John Carroll. 2002. Meaning: A roadmap to knowledge technologies. Em *Proceedings of COLING Workshop A Roadmap for Computational Linguistics*, Taipei, Taiwan.

Rodríguez, Horacio, Salvador Climent, Piek Vossen, Laura Bloksma, Wim Peters, Antonietta Alonge, Francesca Bertagna, e Adriana Roventini. 1998. The top-down strategy for building eurowordnet: Vocabulary coverage, base concepts and top ontology. *Computers and the Humanities*, 32(2-3):117–152.

Rodríguez, Horacio, David Farwell, Javier Farreres, Manuel Bertran, Musa Alkhalifa, Mª Antònia Martí, William J. Black, Sabri Elkateb, James Kirk, Adam Pease, Piek Vossen, e Christiane Fellbaum. 2008. Arabic wordnet: Current state and future extensions. Em *Proceedings of the Fourth International Global WordNet Conference - GWC 2008, Szeged, Hungary*, January, 2008.

Rosch, Eleanor e B. Lloyd. 1978. *Cognition and Categorization*. Lawrence Erlbaum Associates, Hillsdale NJ, USA.

Sagri, Maria Teresa, Daniela Tiscornia, e Francesca Bertagna. 2004. Jur-wordnet. Em *Proceedings of the Second International Global WordNet Conference (GWC'04). Panel on figurative language*, January, 2004.

Sornlertlamvanich, Virach, Thatsanee Charoenporn, e Hitoshi Isahara. 2010. Language resource management system for asian wordnet collaboration and its web service application. Em *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may, 2010. European Language Resources Association (ELRA).

Stamou, Sofia, Alexandros Ntoulas, Jeroen Hoppenbrouwers, Maximiliano Saiz-Noeda, e Dimitris Christoudoulakis. 2002a. Euroterm: Extending the eurowordnet with domain-specific terminology using an expand model approach. Em *Proceedings of the 1st Global WordNet Association conference*, Mysore, India.

Stamou, Sofia, Kemal Oflazer, Karel Pala, Dimitris Christoudoulakis, Dan Cristea, Dan Tufis,

Svetla Koeva, George Totkov, Dominique Dutoit, e Maria Grigoriadou. 2002b. Balkanet: A multilingual semantic network for the balkan languages. Em *Proceedings of the 1st Global WordNet Association conference*.

Stevenson, Mark e Mark A. Greenwood. 2006. Learning Information Extraction Patterns Using WordNet. Em *Proceedings of the 5th Intl. Conf. on Language Resources and Evaluations, LREC 2006 22 - 28 May 2006*, volume 2006, pp. 95–102.

Vossen, Piek. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.

Vossen, Piek, L. Bloksma, Horacio Rodríguez, Salvador Climent, A. Roventini, F. Bertagna, e A. Alonge. 1997. The eurowordnet base concepts and top-ontology. Relatório técnico, Deliverable D017D034D036 EuroWordNet LE2-4003.

Álvez, Javier, Jordi Atserias, Jordi Carrera, Salvador Climent, Egoitz Laparra, Antoni Oliver, e German Rigau. 2008. Complete and consistent annotation of wordnet using the top concept ontology. Em *Proceedings of the the 6th Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech (Morocco), May, 2008.

Álvez, Javier, Paqui Lucio, e G. Rigau. 2012. Adimen-sumo: Reengineering an ontology for first-order reasoning. *International Journal on Semantic Web and Information Systems*, 8(4):80–116.

# Artigos de Investigação

# Un método de análisis de lenguaje tipo SMS para el castellano

## A SMS-like language analyzer for Spanish

Andrés Alfonso Caurcel Díaz
Sistemas inteligentes para la comunicación y movilidad accesibles
Universidad Politécnica de Madrid
AARcaurcel@gmail.com

José María Gómez Hidalgo
Departamento de I+D
Optenet
jgomez@optenet.com

Yovan Iñiguez del Rio
Sistemas inteligentes para la comunicación y movilidad accesibles
Universidad Politécnica de Madrid
yovan.i.rio@gmail.com

**Resumen**

Debido a las características propias del lenguaje tipo SMS utilizado en las comunicaciones por medio de Internet y de los teléfonos móviles, no se puede realizar una tokenización o separación de palabras estándar a la hora de dividir en palabras una oración o frase. La cantidad de elementos no alfanuméricos que se pueden insertar en una palabra, los errores tipográficos y el hecho de no utilizar espacios entre palabras son las principales causas de este problema.

En este artículo presentamos un nuevo sistema de separación de palabras para el análisis del lenguaje natural en español en redes sociales y otras comunicaciones electrónicas. El sistema está integrado en una herramientas para la detección de edad en redes sociales enmarcada en el proyecto de investigación y desarrollo WENDY, y se evalúa cuantitativamente tanto de manera directa, como indirectamente en el marco de dicha aplicación, con resultados positivos en ambos casos.

**Palabras clave**

Lenguaje SMS, lenguaje chat, tokenizador, traductor automático, Procesamiento del Lenguaje Natural, detección de edad

**Abstract**

The usage of specific language codes and chat and SMS-like messages is a major trend in electronic communications. This fact makes Natrual Language Processing quite hard, even at the simplest step fo text message tokenization, due to the widespread usage of non-alphanumeric symbols, frequent typos and non-standard word separators.

In this work we present a new approach for text message tokenization, specific for the Spanish language as used in Social Networks and in electronic communications. Our system has been integrated in a more general application for age-detection in Social Networks developed in the research and development project WENDY, and it has been quantitatively evaluated both in a direct fashion, and indirectly by its impact on the genearl age-detection application, showing very promising results.

**Keywords**

SMS language, chat language, tokenizer, automated translation, Natural Language Processing, Age detection

## 1 Introducción

A la hora de crear aplicaciones informáticas que traten con los nuevo sistemas de comunicación como la mensajería instantánea, chats, foros y redes sociales, es imprescindible abordar la problemática de las características propias del uso del idioma en estos entornos. El uso del idioma en estos ámbitos ha dado lugar a una suerte de dialecto escrito, que con frecuencia se llama lenguaje SMS o lenguaje tipo chat (Forsyth, 2007). Algunas de las características de dicho lenguaje no dependen del idioma del que derivan (por ejemplo español vs. inglés), como el uso de emoticonos, determinadas pautas como la repetición de vocales o la eliminación completa de las mismas, o el uso abusivo de mayúsculas o su ausencia total. Otras características sí que dependen del idioma, proviniendo de siglas de expresiones populares (por ejemplo “LoL” - “Laughing out Loud” en inglés, o “a.p.s.” - “amigos para siempre” en español), o de abreviaciones fonéticas (por ejemplo “cya” - “see you” en inglés, o “xq” - “porque” en español).

Estas características pueden dificultar la lectura a algunos usuarios, y desde luego siempre a cualquier sistema de análisis automático del lenguaje. Por ejemplo, una frase como “felicidadees!! k t lo pases muy bien!! =)Feeeliiciidaaadeeess !! ( :Felicidadesss!!pasatelo genialll :DFeliicCiidaDesS! :D Q tte Lo0 paseS bN! ;) (heart)” puede resultar muy complicada tanto para su lectura por parte de un usuario, como para el análisis por parte de un sistema usualmente preparado para el lenguaje normalizado del que deriva, el español estándar. Incluso la simple fragmentación de la frase anterior en constituyentes básicos como las palabras que la componen,

proceso denominado usualmente tokenización[1], se convierte en un proceso no trivial y cuyos errores pueden llevar a importantes pérdidas de efectividad de cualquier sistema de análisis del lenguaje natural.

Asimismo, el uso de símbolos de puntuación y caracteres no alfanuméricos para símbolos y dibujos obliga a que el tokenizador también respete dichas estructuras para su clasificación y su traducción al lenguaje normalizado si esta fuese necesaria. Los sistemas de tokenización estándar utilizan estos símbolos como límites de palabra, por lo que son eliminados o separados del resto de símbolos (dependiendo de las opciones del tokenizador), perdiéndose por tanto la información que acarrean.

La tokenización de textos electrónicos provenientes de comunicaciones informales en Internet (chats, mensajes cortos, comentarios en foros, etc.) es un tema que ha despertado bastante interés en la comunidad del Procesamiento del Lenguaje Natural, pero los estudios que se han realizado hasta el momento son fundamentalmente para el idioma inglés – por ejemplo (Forsyth, 2007), ya que éste es el más utilizado por los usuarios de la red. No existe ningún estudio comparativo para lenguas romances, aunque existen trabajos para japonés, chino y turco (Ptaszynski, 2010) (Yin, 2009) (Pendar, 2007).

En este trabajo presentamos un nuevo sistema para la tokenización del español en el contexto del lenguaje informal escrito en redes sociales, chats, SMS y mensajería instantánea, orientado a mejorar el reconocimiento de los constituyentes de la oración. Se trata de un sistema que trabaja en dos fases, extrayendo primero los candidatos a constituyentes por medio de una tokenización simple, para luego analizarlos y descomponerlos sucesivamente de acuerdo con una serie de patrones de uso del lenguaje tipo SMS y con la ayuda de recursos lingüísticos. El sistema se dirige exclusivamente hacia el idioma español, y no se ha evaluado su posible validez sobre otros idiomas.

Este tokenizador forma parte de un sistema general orientado a la detección de la edad de los usuarios de redes sociales de acuerdo con patrones de biometría del comportamiento, enmarcado en en el proyecto de investigación de protección al menor WENDY (WEb-access coNfidence for chilDren and Young[2]). Este marco ha permitido evaluar el método de tokenización tanto de manera directa (en función de sus resultados explícitos), como de manera indirecta (en función de como afecta a la efectividad del sistema de detección de edad). En ambos casos se han obtenido resultados positivos.

El resto de este artículo está organizado de la siguiente manera. En primer lugar presentamos el marco general de trabajo, que es el proyecto WENDY, y que permite definir los requisitos particulares del procesamiento de texto en el entorno de las redes sociales. A continuación describimos el sistema de tokenización desarrollado, y seguidamente el entorno de evaluación y los resultados obtenidos. Finalizamos el artículo con las conclusiones obtenidas y las propuestas de trabajo futuro.

## 2 Marco general: el proyecto WENDY

La investigación propuesta en este trabajo se engloba dentro del proyecto de investigación WENDY, que plantea el desarrollo de un sistema de clasificación para detectar la edad de usuarios de redes sociales en castellano relativamente similar al presentado en (Tam, 2009). Los rangos propuestos en el sistema son menores de catorce años (-14) mayores de catorce años pero menores de edad (+14) y mayores de edad (+18). Estos rangos están motivados por la existencia de dos comportamientos significativos a detectar en el marco de la protección del menor en redes sociales:

- Los menores de 14 años no pueden estar dados de alta en una red social sin consentimiento paterno de acuerdo con la legislación española, y con mucha frecuencia mienten con respecto a su edad. Por ejemplo, un estudio reciente revela que se puede estimar en unos 5,6 millones el número de menores de 13 años dados de alta en Facebook[3] en EE.UU.
- Los mayores de 18 años que simulan ser menores de edad pueden estarlo haciendo para establecer contacto con menores, y potencialmente convertirlos en víctimas de acoso sexual ("*grooming*"). Es frecuente encontrarse con casos de este tipo en los medios de comunicación[4].

El sistema desarrollado en WENDY se basa en el análisis del comportamiento de los usuarios dentro de la popular red social Tuenti[5]. En esta red social, los usuarios pueden describir sus gustos (cine, libros, etc.), escribir mensajes de estado, comentarios a los mensajes de sus contactos,

1 La expresión "tokenización" se utiliza con frecuencia en los artículos científicos en español dedicados al análisis del lenguaje natural.

2 http://wendy.optenet.com.

3 http://www.europapress.es/portaltic/socialmedia/noticia-hay-mas-millones-ninos-facebook-si-no-quieren-20120920112441.html.

4 http://www.antena3.com/noticias/sociedad/detenido-entrenador-futbol-infantil-acosar-menores-internet_2012102000066.html.

5 http://www.tuenti.com.

mantener conversaciones por chat y subir o ver fotografías y vídeos. El prototipo desarrollado en WENDY analiza todos los elementos textuales y las fotografías con el fin de determinar la edad del usuario, y avisa al administrador de la red social cuando identifica a un usuario cuyo comportamiento no se corresponde con el habitual de su franja de edad, o simplemente cuando induce que es menor de 14 años o mayor de 18 y no refleja su edad en su perfil.

El sistema integra una serie de módulos especializados por cada tipo de información: gustos del perfil, comentarios, fotografías, etc. En este artículo nos centramos específicamente en dos de los elementos textuales, que son el perfil y los comentarios (sus actualizaciones de estado y las respuestas a otros comentarios).

Para el análisis de estos tipos de información, se ha diseñado un sistema basado en aprendizaje que analiza los elementos textuales y entrena una serie de clasificadores orientados a las franjas de edad descritas anteriormente. Para ello se ha confeccionado una colección de datos obtenidos de perfiles reales de Tuenti compuesta por más de 120.000 perfiles distribuidos de la siguiente manera: Menores de 14 años: 372; de 14 a 17 años: 11.530; más de 18 años: 7.432; sin información sobre la edad: 100.670. Se puede observar que la clase de los menores de 14 años está escasamente representada, lo que dificulta enormemente el entrenamiento de clasificadores efectivos.

Una de las hipótesis de trabajo fundamentales en el proyecto es que el uso del lenguaje tipo SMS puede ser un elemento discriminador en la clasificación por edades. Por ello, se han desarrollado una serie de sistemas orientados al procesamiento de este tipo de lenguaje, que hacen uso del tokenizador descrito en este artículo. A continuación describimos los clasificadores de texto utilizados, así como una serie de sistemas de análisis y normalización del lenguaje tipo chat empleados sobre los textos obtenidos en dicha red social.

### 2.1 Clasificador textual

El clasificador de los textos por edades es un clasificador de texto tradicional – véase (Sebastiani, 2002), en el que los textos:

- Se separan en unidades lingüísticas o términos utilizando el tokenizador que describimos más adelante.
- Los términos obtenidos se traducen y normalizan usando el traductor de lenguaje SMS y el sistema Deflogger descritos en las próximas secciones.
- Se representa cada texto por medio de vectores de pesos de términos en base al modelo del espacio vectorial con pesos de tipo TF.IDF.
- Se seleccionan aquellos términos más predictivos de acuerdo con el criterio de que su Ganancia de Información respecto a las clases constituidas por las franjas de edad sea mayor que cero.
- Se aplica el algoritmo de aprendizaje Bayes Ingenuo, que predice la probabilidad de pertenencia de un texto a una determinada franja de edad en función de las probabilidades condicionadas de pertenen-cia de cada término a dichas franjas.

Se ha utilizado la implementación de Bayes Ingenuo (Naive Bayes) incluida en el paquete de aprendizaje WEKA[6] (Hall et al. 2009), con sus opciones por defecto.

### 2.2 Traductor de lenguaje SMS

Se ha desarrollado un sistema de traducción de los elementos textuales utilizando una serie de recursos lingüísticos preexistentes, a saber: un diccionario de lenguaje tipo SMS para el castellano[7], y un diccionario del idioma castellano[8] (Padró et al., 2010).

El sistema traductor funciona de la siguiente manera:

1. Dada una palabra objetivo (posiblemente una expresión en SMS), se busca la misma en el diccionario de castellano.
2. Si la palabra aparece en el diccionario, se finaliza el proceso. Si la palabra no aparece en el diccionario, entonces se busca en el diccionario de lenguaje SMS.
3. Si la palabra aparece en el diccionario SMS, se selecciona el significado (o traducción) más frecuente o popular. Si la palabra no aparece en el diccionario SMS, se deja como está.

Es preciso resaltar que el diccionario SMS no sólo incluye expresiones coloquiales como “xq” (por “porque”), sino que también contiene una cantidad significativa de emoticonos (por ejemplo “:-)”, etc.). De ahí que se precise que la tokenización sea capaz de mantener y reconocer estos símbolos.

El proceso de traducción se aplica a los textos antes de entrenar un clasificador sobre los mismos.

6 http://www.cs.waikato.ac.nz/ml/weka/.
7 http://www.diccionariosms.com/contenidos/.
8 El incluido en el sistema de análisis lingüístico Freeling: http://nlp.lsi.upc.edu/freeling/.

### 2.3 Analizador tipo Deflogger

La informalidad del lenguaje escrito en foros y redes sociales ha motivado el desarrollo de analizadores específicos capaces de hacer frente a fenómenos como los siguientes:

- Contaminación con abreviaturas típicas del lenguaje SMS (por ejemplo "tkm" por "te quiero mucho", "bss" por "besos", "dsp" por "después",etc.).
- Alternancia de mayúsculas y minúsculas (por ejemplo "AlGuNa LeTrA dE uNa CaNcIoN").
- Repetición de letras (por ejemplo "hoooolllaaaaaa!!!!").
- Omisión de la letra "u" en la las sílabas "que" o "qui", o bien, reemplazo por "k" (por ejemplo "qiero", "kiero").
- Reemplazo de la sílaba "ca" o de la letra "c" por la letra "k" (reemplazo fonético, por ejemplo "kompre kfe").
- Faltas ortográficas intencionales como "soi", "voi", "i" en lugar de "y", etc.

Estos fenómenos y otros se encuentran reconocidos y analizados en la herramienta Deflogger[9], que es un sistema desarrollado para normalizar el lenguaje utilizado en la red social Fotolog[10] y aproximarlo al castellano normal, y que hemos integrado en nuestro sistema. Hemos aplicado el sistema Deflogger a los textos disponibles después de su traducción con el traductor de lenguaje tipo SMS, y antes de entrenar el clasificador.

## 3 Descripción del tokenizador

El sistema de separación de términos debe partir de un texto en lenguaje natural (un comentario de texto, un gusto musical, una lista de películas, etc.) y devolver un conjunto de tokens o unidades textuales básicas compuesto por cadenas de caracteres.

El sistema WENDY ha sido implementado en el lenguaje Java, y se ha desarrollado aprovechando el paquete de aprendizaje automático WEKA, que además del algoritmo de aprendizaje, incluye también métodos de selección de atributos por Ganancia de Información, e incluso un tokenizador simple que separa los textos en palabras de acuerdo a una serie de caracteres configurables por el usuario: espacios, tabuladores, saltos de línea, símbolos de puntuación, etc.

El separador de palabras desarrollado en WENDY parte de la implementación de este tokenizador básico, y aplica dos fases de separación. La primera fase consiste en separar la frase inicialmente por medio de espacios en blanco. Cada parte de frase separada por espacios ofrece tres posibilidades:

- Que sea una palabra propiamente dicha (secuencia de caracteres alfanuméricos).
- Que sea una estructura de signos de puntuación y caracteres no alfanuméricos (posiblemente un emoticono).
- Que sea una mezcla de ambas.

En los dos primeros casos, se considera que las secuencias de caracteres son ya palabras de por sí. Esto implica que las secuencias de caracteres no alfanuméricos, que en particular incluyen a los símbolos de puntuación aislados, se consideran tokens relevantes.

En el tercer caso, se procede a una segunda tokenización de la sub-frase, reconociendo como tokens todas las sub-secuencias de caracteres del mismo tipo – alfanuméricos vs. no alfanuméricos. Por ejemplo, dada la secuencia "Hola:-)ketal", ésta se separaría en esta segunda fase en los tokens siguientes: "Hola", ":-)", y "ketal".

Adicionalmente, se ha tenido en cuenta el sistema de emoticonos propio del chat de Tuenti, en el que las letras en mayúsculas entre paréntesis se representan automáticamente con un símbolo. Por ejemplo, "(L)" se representa gráficamente como un corazón, o "(M)" como una nota musical.

En un ejemplo real, si aplicamos el proceso completo a la frase siguiente:

> *"Felicidades LauraHey, felicidades! ^^felicidiadeees;DFelicidades!Un beso! FELIZIDADESS LAURIIIIIIIIIIII (LL)felicidadeeeeeees! :D jajaja mira mi tablonme meo jajajajajjajate quiero(:,"*

Se obtiene como salida un vector cuyos componentes se muestran en la tabla 1:

| | | |
|---|---|---|
| *Felicidades* | *LauraHey* | *,* |
| *felicidades* | *!* | *^^* |
| *felicidiadeees* | *;* | *DFelicidades* |
| *!* | *Un* | *beso* |
| *!* | *FELIZIDADESS* | *LAURIIIIIIIIIIII* |
| *(LL)* | *felicidadeeeeeees* | *!* |
| *:D* | *jajaja* | *mira* |
| *mi* | *tablonme* | *meo* |
| *jajajajajjajate* | *quiero* | *(:,* |

Tabla 1: Vector de salida de ejemplo usando el nuevo sistema de separación.

9 http://code.google.com/p/deflog/.

10 http://www.fotolog.com.

Como se puede observar, la tokenización no es perfecta. Por ejemplo, en la expresión "felicidiadeees;DFelicidades" no se reconoce individualmente como emoticono ";D", debido a que la letra "D" se considera anexa a la siguiente palabra. Igualmente, no se separan las palabras "jajajajajjaja" y "te", ya que se trata de una única secuencia alfanumérica.

La implementación del tokenizador ha sido realizada en Java, partiendo del tokenizador previamente existente `NgramTokenizer`[11] ya existente en el paquete de aprendizaje WEKA. Este sistema de separación genera secuencias de tokens de las longitudes deseadas (ngramas de tokens), y acepta los siguientes parámetros:

- Secuencia de separadores, que es una cadena que incluye los elementos que se usan como separadores entre tokens. La cadena de separadores por defecto es: " \r\n\t.,;:'"()?!".
- Número mínimo de tokens en secuencia, por defecto 1.
- Número máximo de tokens en secuencia, por defecto 3.

Los últimos dos parámetros determinan el tamaño N de los ngramas. Este tokenizador se utiliza de manera integrada dentro del filtro `StringToWordVector`[12] incluido en el paquete WEKA, que acepta un tokenizador o separador como parámetro entre otros, y que transforma una colección de ejemplares textuales en una representación de bolsa de palabras o de vectores de pesos de términos según el Modelo del Espacio Vectorial (Sebastiani, 2002).

Dado que el nuevo separador desarrollado, que hemos llamado `SMSTokenizer`, desciende de la clase `Tokenizer` del paquete WEKA, puede ser usado como cualquier otro separador dentro de este paquete de aprendizaje.

## 4 Evaluación del tokenizador

Para evaluar la utilidad de este nuevo método de separación, nos hemos planteado dos preguntas importantes:

- ¿Se identifican mejor los tokens del lenguaje?
- ¿El nuevo tokenizador mejora los resultados de clasificadores en textos SMS?

La primera pregunta se corresponde con una evaluación directa o de tipo caja de cristal, donde se analizan los resultados concretos del proceso de tokenizado. La segunda pregunta se corresponde con una evaluación indirecta o de tipo caja negra, y es también la más importante por cuanto el objetivo del sistema global de WENDY es reconocer con la máxima eficacia posible la edad de los usuarios en función de sus interacciones textuales.

En ambos casos se ha partido del conjunto de frases de comentarios descargados de Tuenti con edades identificadas en sus perfiles, obteniéndose un total de 84.956 frases. Los textos tienen un alto contenido de lenguaje SMS. En dichos textos, han sido traducidas las marcas de lenguaje HTML que contenían y se han sustituido las marcas propias de la red social por medio de identificadores de las mismas (e.g: hiperenlaces, frases automáticas del sistema). La distribución de comentarios por rango de edad se muestra en la Tabla 2.

| Clase | Número | Porcentaje |
|---|---|---|
| -14 | 1.810 | 2,13% |
| +14 | 59.778 | 70,36% |
| +18 | 23.368 | 27,51% |
| Total | 84.956 | 100,00% |

Tabla 2: Distribución de comentarios por clase.

Como se puede observar, y en línea con las estadísticas generales de la colección de datos descritas en la Sección 2, la clase más sensible (los menores de 14 años, denotada por "-14") es la menos representada, y la siguiente clase más sensible (los mayores de 18 años, denotados por "+18") es bastante minoritaria. Dada esta distribución de clases, podemos afirmar que la detección de la edad vinculada con estos comentarios es un problema bastante difícil, ya que los algoritmos de aprendizaje tienden a intentar optimizar la eficacia global y no la eficacia sobre determinadas clases. Por ejemplo, en este caso un sistema trivial que clasificase todo comentario como perteneciente a un menor entre 14 y 17 años (clase denotada por "+14"), tendría ya una eficacia del 70%.

Como sistema de tokenización de referencia o línea base, hemos utilizado el NgramTokenizer original, centrándonos en los ngramas de longitud uno (es decir, se invoca con número mínimo y máximo de tokens 1, y con los separadores por defecto).

### 4.1 Evaluación directa

Para evaluar de manera directa la calidad del nuevo sistema, debemos tener en cuenta dos aspectos:

- En primer lugar, el objetivo es reconocer mejor los elementos individuales del lenguaje o palabras aisladas, ya sean palabras del castellano estándar o palabras incluidas en el

11 http://weka.sourceforge.net/doc.dev/weka/core/tokenizers/NGramTokenizer.html.

12 http://weka.sourceforge.net/doc.dev/weka/filters/unsupervised/attribute/StringToWordVector.html.

diccionario de lenguaje SMS usado en el proyecto WENDY.

- En segundo lugar, el procesamiento de los textos no es un proceso aislado, sino que está enmarcado dentro del proceso general de detección de edad, por lo que a los textos originales se le debe aplicar el procedimiento antes mencionado de traducción de lenguaje tipo SMS.

Recordando que el proceso de traducción de lenguaje tipo SMS busca secuencialmente las palabras o tokens aislados en los diccionarios del castellano estándar y en el diccionario SMS, se ha realizado el proceso de traducción y se han computado el número de palabras detectadas sucesivamente en cada uno de los pasos de la traducción. A continuación, hemos comparado las palabras encontradas en cada diccionario usando el separador por defecto y el nuevo método de separación.

En las tablas 3, 4 y 5 se muestra la diferencia entre el número de tokens de cada comentario encontrado en el diccionario de castellano estándar (tabla 3), en el diccionario de lenguaje SMS (tabla 4) y no encontradas (tabla 5). En la primera columna se muestra el número medio de palabras de diferencia entre usar el tokenizador original y el nuevo tokenizador, mientras que en la segunda columna se muestra la desviación típica de esta diferencia, y en la tercera y cuarta columnas se muestran las diferencias mínimas y máximas.

Por ejemplo, examinando la tabla 3, se puede observar que el nuevo tokenizador reconoce 9,5 palabras más que el tokenizador básico por comentario, en media. Es decir, que si al realizar el proceso de traducción se encuentran X palabras de un comentario en el diccionario de castellano estándar usando el separador original, usando el nuevo separador se encuentran X + 9,5, lo que indica que el nuevo tokenizador reconoce mejor las palabras del castellano estándar incluidas dentro de los comentarios.

En la tabla 4 también se puede observar una cierta ganancia (en este caso, sólo de 1,13 palabras de media reconocidas por comentario), mientras que en la tercera tabla, el comportamiento es coherente, ya que al mostrar un número negativo (que es igual a la suma de los correspondientes a las tablas 3 y 4), indica que quedan menos palabras o tokens por reconocer por cada comentario. Cabe señalar que es razonable esperar que se localicen menos términos en el diccionario SMS, ya que sólo se buscan en él los términos que no han sido localizados en el diccionario de castellano, y que su tamaño es de dos órdenes de magnitud menor que el del diccionario del castellano.

| Media | Desviación | Mínimo | Máximo |
|---|---|---|---|
| 9,5 | 14,63 | -58,93 | 100 |

Tabla 3: Diferencia de aciertos en diccionario castellano.

| Media | Desviación | Mínimo | Máximo |
|---|---|---|---|
| 1,13 | 4,09 | -41,67 | 50 |

Tabla 4: Diferencia de aciertos en elementos SMS.

| Media | Desviación | Mínimo | Máximo |
|---|---|---|---|
| -10,62 | 15,56 | -100 | 58,93 |

Tabla 5: Diferencias de tokens no encontrados.

Dado que existe una desviación típica tan grande en las tabla 3 y 4, hemos tomado la decisión de realizar un análisis más detallado teniendo en cuenta el tamaño de los distintos comentarios.

| Rango | Tokens totales | Aciertos castellano | Aciertos SMS | Fallos |
|---|---|---|---|---|
| 1 – 6 | 19.345 | 4.625 | 298 | 14.422 |
| 7 – 20 | 169.246 | 82.966 | 5.379 | 88.967 |
| 21 – 50 | 403.366 | 210.916 | 13.469 | 178.981 |
| 51 – 99 | 519.784 | 270.146 | 18.982 | 230.656 |
| 100 – 250 | 1,247.541 | 656.809 | 44.935 | 545.797 |

Tabla 6: Frecuencias por franjas con el tokenizador básico.

| Rango | Tokens totales | Aciertos castellano | Aciertos SMS | Fallos |
|---|---|---|---|---|
| 1 – 6 | 29.322 | 12.793 | 1.065 | 15.464 |
| 7 – 20 | 258.109 | 142.806 | 11.033 | 104.270 |
| 21 – 50 | 601.453 | 344.710 | 25.361 | 231.382 |
| 51 – 99 | 802.203 | 448.135 | 35.095 | 318.973 |
| 100 – 250 | 1,939.077 | 1,081.811 | 83.002 | 774.264 |

Tabla 7: Frecuencias por franjas con el nuevo tokenizador.

Presentamos este análisis en las tablas 6 y 7, mostrando en la primera de ellas el comportamiento del tokenizador básico, mientras que en la segunda mostramos el comportamiento del nuevo sistema de tokenización presentado en este artículo. En cada tabla mostramos por fila los resultados para los comentarios con una longitud en número de palabras dentro de una serie de rangos (de 1 a 16 palabras, de 7 a 20, etc.). En cada columna mostramos el número de palabras totales localizadas, cuántas de ellas se encuentran en el diccionario en castellano, cuántas en el diccionario SMS y cuántas

no se encuentran en ningún diccionario (fallos de búsqueda).

La primera observación a realizar al comparar ambas tablas es que las celdas de la segunda están pobladas con cantidades mucho mayores que la primera – por ejemplo, observando la primera columna en ambas tablas, se concluye que el número de tokens detectado en la segunda (por el nuevo tokenizador) es superior aproximadamente en un 50% al obtenido por el tokenizador básico. Con las demás columnas se puede hacer un análisis similar, siendo los incrementos próximos al 60% en ocasiones. Sin embargo, este hecho no es factor relevante desde el punto de vista del análisis de la calidad en la separación de palabras, ya que debemos tener en cuenta que el nuevo tokenizador admite como tokens las secuencias de caracteres no alfanuméricos (potenciales emoticonos), mientras que el primero considera la gran mayoría de signos de puntuación como separadores y por tanto no como tokens.

El análisis de los resultados de las tablas 6 y 7 debe ser relativo, es decir: ¿aumenta el número de tokens encontrados en los diccionarios en términos relativos (sobre el número de tokens encontrados) cuando se compara ambos tokenizadores? De ser así, la separación de palabras realizada por el nuevo tokenizador sería de mayor calidad.

Para que sea más sencillo analizar los resultados, mostramos también las 8 y 9, en las que se muestran los porcentajes de elementos correspondientes a cada una de las columnas anteriores.

| Rango | % castellano | % SMS | % fallos |
|---|---|---|---|
| 1 – 6 | 23,91% | 1,54% | 74,55% |
| 7 – 20 | 49,02% | 3,18% | 52,57% |
| 21 – 50 | 52,29% | 3,34% | 44,37% |
| 51 – 99 | 51,97% | 3,65% | 44,38% |
| 100 – 250 | 52,65% | 3,60% | 43,75% |
| Media | 45,97% | 3,06% | 51,92% |

Tabla 8: Porcentajes correspondientes a las frecuencias obtenidas con el tokenizador básico.

| Rango | % castellano | % SMS | % fallos |
|---|---|---|---|
| 1 – 6 | 43,63% | 3,63% | 52,74% |
| 7 – 20 | 55,33% | 4,27% | 40,40% |
| 21 – 50 | 57,31% | 4,22% | 38,47% |
| 51 – 99 | 55,86% | 4,37% | 39,76% |
| 100 – 250 | 55,79% | 4,28% | 39,93% |
| Media | 53,58% | 4,16% | 42,26% |

Tabla 9: Porcentajes correspondientes a las frecuencias obtenidas con el nuevo tokenizador.

Con carácter global (en media), con el nuevo tokenizador se reconocen un 53,58% de los tokens en el diccionario de castellano mientras que con el tokenizador básico, el porcentaje es del 45,97%. La mejora en el reconocimiento dentro del diccionario SMS es menor pero relevante, ya que pasamos de un 3,06% a un 4,16%, y en términos de tokens no reconocidos, éstos descienden casi en un 10%.

El comportamiento es similar en ambas tablas, ya que el porcentaje de elementos localizados en los diccionarios tanto cuando se utiliza el tokenizador básico como cuando cuando se utiliza el nuevo tokenizador, es mayor en los rangos centrales, mientras que es menor en los rangos extremos. Existe una diferencia mayor particularmente entre el rango 1 a 6 y los demás rangos, y es precisamente en ese rango donde se produce una mayor mejora en los porcentajes de tokens reconocidos en los diccionarios, y en consecuencia, en los tokens no reconocidos o fallos, que se reducen drásticamente de un 74,55% a un 52,74%.

Por contraste, en los comentarios más largos la mejora es la menor, lo cual se debe probablemente a que, al tratarse de textos más largos, los usuarios ponen una mayor atención en su redacción – en otras palabras, si se toman su tiempo en escribir un comentario de 250 palabras, también se molestan en escribirlo en un castellano más correcto.

### 4.2 Evaluación indirecta

La evaluación indirecta del nuevo tokenizador consiste en examinar su impacto en la tarea objetivo, que es el reconocimiento de edad según se ha descrito anteriormente. Para ello, se construyen clasificadores siguiendo los pasos descritos en la sección 2.1, usando el tokenizador básico y el nuevo tokenizador, para comparar sus resultados en términos de eficacia.

El proceso consiste en la tokenización de los 84.956 comentarios, la traducción SMS y la normalización con Deflogger de los términos obtenidos, la representación de los comentarios como vectores de pesos de términos, el filtrado de los términos por Ganancia de Información, y el entrenamiento de un clasificador Bayes Ingenuo (Naive Bayes[13]) incluido en WEKA con sus opciones por defecto.

Cabe resaltar que el tokenizador estándar es capaz de localizar 227.781 términos distintos, mientras que el nuevo tokenizador detecta 272,769 términos diferentes (incluyendo por ejemplo las secuencias de símbolos de puntuación, interpretadas como emoticonos). Sin embargo, una vez se aplica el filtro de términos por su valor de Ganancia de

13 http://weka.sourceforge.net/doc/weka/classifiers/bayes/NaiveBayes.html.

Información, en el clasificador basado en el tokenizador básico quedan 38.564 términos, mientras que en el basado en el nuevo tokenizador quedan 26.233, es decir, muchos menos. Nosotros interpretamos esta circunstancia como una mejora, ya que el nuevo tokenizador aumenta la calidad de los términos usados en la representación al concentrar sus estadísticas – en otras palabras, términos que anteriormente eran poco frecuentes y aportaban información positiva pero marginal, pasan a estar concentrados en términos más frecuentes que aportan más información al tener frecuencias mayores.

Para evaluar los clasificadores obtenidos, se ha aplicado un proceso de validación cruzada en cuatro carpetas (es decir, se han ejecutado cuatro experimentos usando en cada uno de ellos, un 75% de la colección de datos para entrenamiento, y un 25% para evaluación). Como métricas de evaluación, y teniendo en cuenta que es preciso observar el detalle de las clases más sensibles (menores de 14 años y mayores de 18 años), se ha calculado la tabla de confusión en cada caso (agregada sobre los 4 experimentos) y la precisión y la cobertura por cada clase.

En las tablas 10 y 11 se muestran los resultados obtenidos por los clasificadores entrenados usando el tokenizador básico y el nuevo tokenizador, respectivamente. En cada fila se muestran las poblaciones de las clases, mientras que por columnas se muestran las decisiones de clasificación. Por tanto, las cuatro primeras filas y columnas constituyen la tabla de confusión. Los aciertos aparecen en la diagonal principal de cada tabla. En las dos últimas columnas se muestran la precisión y la cobertura por cada clase.

| | -14 | 14 | 18 | Precisión | Cobertura |
|---|---|---|---|---|---|
| -14 | 530 | 1232 | 48 | 0,61 | 0,29 |
| 14 | 262 | 57298 | 2218 | 0,77 | 0,95 |
| 18 | 73 | 15850 | 7445 | 0,76 | 0,31 |

Tabla 10: Tabla de confusión y valores de efectividad con el tokenizador por defecto.

| | -14 | 14 | 18 | Precisión | Cobertura |
|---|---|---|---|---|---|
| -14 | 597 | 1146 | 67 | 0,63 | 0,32 |
| 14 | 290 | 57336 | 2152 | 0,77 | 0,96 |
| 18 | 69 | 15752 | 7546 | 0,77 | 0,32 |

Tabla 11: Tabla de confusión y valores de efectividad con el nuevo tokenizador propuesto.

Como se puede observar, el proceso de clasificación arroja mejores resultados con el nuevo proceso de tokenización. En números globales, el nuevo tokenizador conlleva un incremento de efectividad del 0,768 al 0,770, que se corresponde con 206 nuevos aciertos. Aunque en términos absolutos la mejora puede no parecer muy significativa, esta mejora se concentra sobre todo en las clases menos representadas y más significativas, que son las de menores de 14 años y las de mayores de 18, por lo que a efectos prácticos desde el punto de vista de la eficacia de la detección de edad, consideramos que la mejora es importante.

## 5 Conclusiones y trabajo futuro

En este artículo hemos presentado un nuevo sistema de tokenización específico para el texto informal que utilizan los usuarios de redes sociales, enmarcado en un sistema de detección de edad en este contexto. El sistema de tokenización ha sido evaluado con resultados positivos, tanto desde el punto de vista de la calidad explícita de los términos detectados, como desde el del impacto que tiene en la efectividad del clasificador de edad.

En futuros trabajos planeamos evaluar el sistema de tokenización no sólo con el algoritmo de aprendizaje propuesto, sino también con otros algoritmos de aprendizaje, integrandolo en el sistema multimodal de reconocimiento de edad que utiliza otros elementos de información (perfiles, fotografías, etc.).

## Agradecimientos

Esta investigación ha sido financiada por Optenet S.A. y por el Ministerio de Economía y Competitividad y el Centro para el Desarrollo Tecnológico Industrial (CDTI), en el marco del proyecto de investigación industrial "*WENDY: WEb-access coNfidence for chilDren and Young*" (TSI-020100-2010-452).

## Referencias

Forsyth Eric N. and Craig H. Martell, *Lexical and Discourse Analysis of Online Chat Dialog,* Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007), pp. 19-26, September 2007.

Hall Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten (2009); *The WEKA Data Mining Software: An Update*; SIGKDD Explorations, Volume 11, Issue 1.

Padró, Lluís, Miquel Collado, Samuel Reese, Marina Lloberes and Irene Castellón. *FreeLing 2.1: Five Years of Open-Source Language Processing Tools.* Proceedings of 7th Language

Resources and Evaluation Conference (LREC 2010), ELRA.
La Valletta, Malta. May, 2010.

Pendar, Nick Nick Pendar, *Toward Spotting the Pedophile Telling victim from predator in text chats*, 2012 IEEE Sixth International Conference on Semantic Computing, pp. 235-241, International Conference on Semantic Computing (ICSC 2007), 2007

Ptaszynski, Michal, Pawel Dybala, Tatsuaki Matsuba, Fumito Masui, Rafal Rzepka and Kenji Araki, *Machine Learning and Affect Analysis Against Cyber-Bullying*, In Proceedings of The Thirty Sixth Annual Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB'10), 29th March – 1st April 2010, De Montfort University, Leicester, UK, pp. 7-16, 2010.

Sebastiani, F. 2002. *Machine learning in automated text categorization*. ACM Computing Surveys, 34(1):1–47.

Tam, J., and C. H. Martell. 2009. *Age detection in chat.* Paper presented at Semantic Computing, 2009. ICSC '09. IEEE International Conference on, .

Yin D., Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, *Detection of Harassment on Web 2.0* in CAW 2.0 '09: Proceedings of the 1st Content Analysis in Web 2.0 Workshop, Madrid, Spain, 2009.

# Extracção de Relações Semânticas de Textos em Português Explorando a DBpédia e a Wikipédia

**Exploring DBpedia and Wikipedia for Portuguese Semantic Relationship Extraction**

David S. Batista David Forte Rui Silva Bruno Martins
Mário J. Silva
Instituto Superior Técnico and INESC-ID, Lisboa, Portugal
{david.batista,david.forte,rui.silva,bruno.g.martins,mjs}@ist.utl.pt

**Resumo**

A identificação de relações semânticas, expressas entre entidades mencionadas em textos, é um passo importante para a extracção automática de conhecimento a partir de grandes colecções de documentos, tais como a Web. Vários trabalhos anteriores abordaram esta tarefa para o caso da língua inglesa, usando técnicas de aprendizagem automática supervisionada para classificação de relações, sendo que o actual estado da arte recorre a métodos baseados em *string kernels* (Kim et al., 2010; Zhao e Grishman, 2005). No entanto, estas abordagens requerem dados de treino anotados manualmente para cada tipo de relação, além de que os mesmos têm problemas de escalabilidade para as dezenas ou centenas de diferentes tipos de relações que podem ser expressas. Este artigo discute uma abordagem com supervisão distante (Mintz et al., 2009) para a extracção de relações de textos escritos em português, a qual usa uma técnica eficiente para a medição de similaridade entre exemplares de relações, baseada em valores mínimos de dispersão (i.e., *min-hashing*) (Broder, 1997) e em dispersão sensível à localização (i.e., *Locality-Sensitive Hashing*) (Rajaraman e Ullman, 2011).

No método proposto, os exemplos de treino são recolhidos automaticamente da Wikipédia, correspondendo a frases que expressam relações entre pares de entidades extraídas da DBPédia. Estes exemplos são representados como conjuntos de tetragramas de caracteres e de outros elementos representativos, sendo os conjuntos indexados numa estrutura de dados que implementa a ideia da dispersão sensível à localização. Procuram-se os exemplos de treino mais similares para verificar qual a relação semântica que se encontra expressa entre um determinado par de entidades numa frase, com base numa aproximação ao coeficiente de Jaccard obtida por *min-hashing*. A relação é atribuída por votação ponderada, com base nestes exemplos. Testes com um conjunto de dados da Wikipédia comprovam a adequabilidade do método proposto, tendo sido extraídos 10 tipos diferentes de relações, 8 deles assimétricos, com uma pontuação média de 55.6% em termos da medida $F_1$.

**Palavras chave**

Extracção de Relações, Extracção de Informação

**Abstract**

The identification of semantic relationships, as expressed between named entities in text, is an important step for extracting knowledge from large document collections, such as the Web. Previous works have addressed this task for the English language through supervised learning techniques for automatic classification. The current state of the art involves the use of learning methods based on *string kernels* (Kim et al., 2010; Zhao e Grishman, 2005). However, such approaches require manually annotated training data for each type of semantic relationship, and have scalability problems when tens or hundreds of different types of relationships have to be extracted. This article discusses an approach for distantly supervised relation extraction over texts written in the Portuguese language, which uses an efficient technique for measuring similarity between relation instances, based on minwise hashing (Broder, 1997) and on locality sensitive hashing (Rajaraman e Ullman, 2011).

In the proposed method, the training examples are automatically collected from Wikipedia, corresponding to sentences that express semantic relationships between pairs of entities extracted from DBPedia. These examples are represented as sets of character quadgrams and other representative elements. The sets are indexed in a data structure that implements the idea of locality-sensitive hashing. To check which semantic relationship is expressed between a given pair of entities referenced in a sentence, the most similar training examples are searched, based on an approximation to the Jaccard coefficient, obtained through min-hashing. The relation class is assigned with basis on the weighted votes of the most similar examples. Tests with a dataset from Wikipedia validate the suitability of the proposed method, showing, for instance, that the method is able to extract 10 different types of semantic relations, 8 of them corresponding to asymmetric relations, with an average score of 55.6%, measured in terms of $F_1$.

**Keywords**

Relation Extraction, Information Extraction

## 1 Introdução

Em Extracção de Informação (EI) e Processamento de Linguagem Natural (PLN), a tarefa de extracção de relações consiste em detectar e classificar relações semânticas, em colecções de documentos textuais. Por exemplo, na frase ***Brooklyn*** *é um dos 62 condados do estado americano de* ***Nova Iorque***, a relação semântica *localizado-em* encontra-se expressa entre dois nomes de locais. Alguns domínios de aplicação incluem a detecção de relações de interação entre pares de proteínas, ou entre genes e doenças, na literatura biomédica (Bunescu e Mooney, 2005a; Kim et al., 2010; Zhou e Zhang, 2007), a detecção de diferentes tipos de associações entre entidades mencionadas em textos jornalísticos, tais como as relações entre pessoas e os seus locais de nascimento, ou entre pessoas e as organizações a que pertencem (Hachey, Grover e Tobin, 2012), ou ainda a detecção de relações semânticas entre pares de expressões nominais em geral (Hendrickx et al., 2010). Ao longo dos anos, têm sido propostas diferentes abordagens para resolver a tarefa de extracção de relações. Em particular, os métodos baseados em regras aplicam regras linguísticas para capturar os padrões tipicamente usados para expressar relações (Brin, 1999). Os métodos baseados em características, por outro lado, transformam exemplos das relações semânticas a extrair em conjuntos de características linguísticas, como por exemplo características lexicais, sintácticas e/ou semânticas, capturando a semelhança entre vectores de características através de técnicas de aprendizagem automática supervisionada (GuoDong et al., 2005). Os trabalhos mais recentes na área envolvem a utilização de métodos de aprendizagem baseados em *string kernels*, quer explorando *kernels* para representar sequências (Bunescu e Mooney, 2005a), numa tentativa de capturar padrões sequenciais dentro de frases dos textos, ou *kernels* específicos para árvores ou para grafos no geral, por forma a aprender funções de classificação relacionadas com os padrões em estruturas resultantes de uma análise sintáctica (Nguyen, Moschitti e Riccardi, 2009; Bunescu e Mooney, 2005b). Os métodos baseados em *string kernels* são superiores aos métodos baseados em características, no sentido de melhor contornarem o facto de os dados de treino serem tipicamente muito esparsos, ou no sentido de proporcionarem uma exploração eficiente de espaços de características muito grandes. Porém, estes métodos são computacionalmente exigentes quando se considera um número elevado de classes, ou sempre que é preciso manipular conjuntos de treino grandes, tornando assim difícil a sua aplicação em problemas reais.

Neste trabalho, propomos explorar uma abordagem diferente para a extracção automática de relações semânticas, com base na pesquisa pelos $kNN$ exemplos de treino mais próximos, como forma de fazer a classificação, aproveitando um método eficiente baseado em valores mínimos de funções de dispersão como forma de medir a similaridade entre relações, para diferentes tipos de relações semânticas. O método proposto é avaliado na tarefa específica de extracção de relações em textos escritos em português, sendo os exemplos de treino extraídos automaticamente da Wikipédia, com base nas relações entre pares de entidades que se encontram explicitamente codificados na DBPédia. Desta forma, através de supervisão distante (Hoffmann, Zhang e Weld, 2010), contornamos a dificuldade em coleccionar exemplos de treino anotados.

Num trabalho anterior explorámos já a ideia de pesquisar pelos $kNN$ exemplos de treino mais próximos, como forma de abordar a extracção de relações de textos em inglês, de forma supervisionada e usando conjuntos de dados bem conhecidos na área (Batista et al., 2013). Com este trabalho pretendemos agora testar a eficiência desta abordagem num cenário envolvendo supervisão distante e textos em português.

Desta forma, realizámos experiências exaustivas com diferentes configurações do método de classificação proposto, variando o tamanho das assinaturas de *min-hash*, assim como o número de exemplos mais próximos considerados para a classificação. Experiências com um conjunto de dados da Wikipédia portuguesa comprovam a adequabilidade do método proposto. Os melhores resultados correspondem a uma macromédia de 55.6% em termos da medida $F_1$, quando se consideram 10 tipos diferentes de relações semânticas, 8 delas correspondendo a relações semânticas assimétricas.

O resto deste artigo está organizado da seguinte forma: a Secção 2 apresenta trabalhos relacionados importantes. A Secção 3 descreve o método proposto, detalhando a recolha automática de exemplos de treino a partir da Wikipédia e da DBPédia, descrevendo a representação considerada para os exemplos dos diferentes tipos de relações, e apresentando a abordagem de dispersão sensível à localização. A Secção 4 apresenta a avaliação experimental do método proposto. Finalmente, a Secção 5 resume as nossas principais conclusões e apresenta orientações possíveis para trabalho futuro.

## 2 Trabalho Relacionado

Extrair relações semânticas entre expressões nominais (por exemplo, entre nomes de entidades como pessoas, locais ou organizações), tal como mencionadas nos textos, é um passo crucial na compreensão da linguagem natural, com muitas aplicações práticas. Vários autores têm proposto técnicas de aprendizagem automática para abordar o problema, por exemplo formulando o mesmo como uma tarefa de classificação binária (i.e., uma tarefa de classificação supervisionada binária, definida sobre exemplares de candidatos a relações entre pares de expressões nominais, em que os exemplares são classificados como membros de uma classe *exemplares_relacionados* ou de uma classe *exemplares_não_relacionados*).

Apesar do exemplo anterior se focar no caso de relações binárias entre pares de expressões nominais, a discussão é facilmente estendida à extracção de relações considerando $n$ tipos diferentes de relações semânticas entre entidades.

Entre as abordagens anteriores relevantes incluem-se os trabalhos de autores que adoptaram métodos baseados em vectores de características e aprendizagem supervisionada, ou métodos baseados em *string kernels*. A maior vantagem de métodos baseados em *string kernels* reside no facto destas soluções permitirem explorar um espaço de características muito grande (i.e., frequentemente exponencial ou, nalguns casos, infinito) em tempo computacional polinomial, sem a necessidade de representar explicitamente os vectores de características. No entanto, os métodos de classificação baseados em *string kernels* são apesar de tudo muito exigentes em termos de requisitos computacionais, perante problemas que envolvam um número elevado de classes ou grandes colecções de dados.

Dado um conjunto de exemplos positivos e negativos sobre um dado tipo de relação semântica, os métodos baseados em características começam por extrair características sintácticas e semânticas dos textos, utilizando-as como pistas para decidir se as entidades numa dada frase se encontram relacionadas ou não. As características sintácticas extraídas das frases incluem tipicamente:

1. As próprias entidades em si.
2. A categoria semântica das duas entidades (e.g. pessoa, local ou organização).
3. A sequência de palavras entre as entidades.
4. O número de palavras entre as entidades.
5. O caminho entre as duas entidades numa árvore de análise sintáctica (i.e., *parse tree*).

As características semânticas podem incluir, por exemplo, o caminho entre as duas entidades numa estrutura resultante de uma análise de dependências entre os constituintes da frase. Muito embora a análise de dependências possa ser vista como puramente sintáctica, muitos trabalhos anteriores nesta área argumentam que a mesma está próxima de uma representação semântica.

As várias características são apresentadas a um classificador sob a forma de um vector de características. Vários algoritmos de aprendizagem supervisionada, como máquinas de vectores de suporte ou modelos baseados em regressão logística, assim como conjuntos de características diferentes, têm sido explorados na literatura (Zhou e Zhang, 2007; Kambhatla, 2004).

Os métodos baseados em características têm a limitação de envolverem escolhas heurísticas, sendo muitas vezes necessário proceder à selecção de características numa base de tentativa e erro, por forma a maximizar o desempenho. Para solucionar o problema da selecção automática de um conjunto adequado de características, foram desenvolvidos *kernels* especializados para a tarefa de extracção de relações, tirando partido de representações ricas dos exemplos de treino, e explorando estas representações ricas de uma forma exaustiva e implícita.

Bunescu e Mooney (2005a) apresentaram um *kernel* de subsequências, que trabalha com representações dos exemplos baseadas em sequências esparsas, combinando palavras e etiquetas morfológicas (i.e., *POS tags*) por forma a capturar as palavras no contexto em torno das expressões nominais envolvidas nas relações. Três *kernels* de subsequências são usados para calcular a similaridade entre os exemplares, isto é, entre as instâncias de uma dada relação, ao nível das palavras, ou seja, comparando as sequências de palavras que ocorrem (i) antes e entre, (ii) entre, e (iii) entre e depois das expressões nominais envolvidas nas relações. Um *kernel* combinado é então produzido pela soma dos três *sub-kernels* anteriores. Os autores avaliaram a sua abordagem na tarefa de extrair interações proteicas no corpus AImed (i.e., um conjunto de textos constituído por resumos de publicações no MEDLINE[1]), concluindo que os *kernels* de subsequências, em conjunto com classificadores baseados em máquinas de vectores de suporte, melhoram a qualidade dos resultados, em comparação com um sistema baseado em regras. Além disso, Bunescu e Mo-

[1] `http://ipubmed.ics.uci.edu/`

oney (2005a) também argumentaram que representações mais ricas para as palavras no contexto, com base nas etiquetas morfológicas das palavras e nos tipos de entidades envolvidos nas relações, podem levar a resultados melhores com o *kernel* de subsequências do que com uma abordagem baseada num *kernel* de árvores de dependências, tal como proposto por Culotta e Sorensen (2004).

Zelenko, Aone e Richardella (2003) descreveram uma abordagem para extracção de relações que usa um *kernel* para a comparação de representações das frases baseadas em estruturas sintácticas pouco profundas. O *kernel* é desenhado para calcular a similaridade entre duas árvores de análise sintáctica (i.e., *shallow parse trees*), aumentadas com as entidades sob as quais incide a relação, em termos da soma ponderada do número de sub-árvores que são comuns entre duas representações. Estes autores avaliaram a sua abordagem numa tarefa de extracção de relações dos tipos *pessoa-filiação* e *organização-localização*, notando no entanto que o método proposto é vulnerável a erros na geração das árvores de análise sintáctica.

Culotta e Sorensen (2004) descreveram ainda uma versão modificada do *kernel* anterior utilizando representações baseadas em árvores de dependências, em que um *kernel* para a comparação de conjuntos de palavras também é usado como forma de compensar erros na análise sintáctica. Na proposta de Culotta e Sorensen, cada nó da árvore de dependências contém informação rica, como a identidade da palavra, a etiqueta morfológica, o tipo de sintagma (nominal, verbal, etc.), ou o tipo de entidade. Usar uma representação estruturada mais rica pode levar a um ganho de desempenho, em comparação com a utilização de abordagens baseadas simplesmente em sacos de palavras (i.e., *bags of words*). Uma versão refinada foi posteriormente proposta por Zhao e Grishman (2005), usando nós compostos para integrar informações de diferentes fontes sintácticas (por exemplo, informação lexical, informação resultante de uma análise sintáctica, e informação resultante de uma análise de dependências). Desta forma, os erros de processamento que ocorram ao nível das representações podem ser superados por informações provenientes de outros níveis.

Airola et al. (2008) introduziram um *kernel* denominado de *all-dependency-paths*, usando uma representação baseada num grafo direccionado com arestas ponderadas, que combinam dois subgrafos desconexos. Temos assim que, uma estrutura representa a árvore de dependências de uma frase, e uma outra estrutura representa a ordem sequencial das palavras. Bunescu e Mooney (2005) apresentaram ainda uma abordagem alternativa que utiliza a informação concentrada no caminho mais curto na árvore de dependências entre as duas entidades. Estes autores argumentam que o caminho mais curto, entre as duas expressões nominais, numa árvore resultante da análise de dependências, codifica informação suficiente para extrair relações.

Estudos recentes continuam a explorar combinações ou extensões dos métodos baseados em *kernels* descritos anteriormente (Kim et al., 2010). No entanto, a maioria das abordagens propostas são avaliadas em conjuntos de dados diferentes, pelo que não é possível ter uma ideia clara sobre qual a abordagem que é efectivamente melhor. Os conjuntos de dados habitualmente usados na língua inglesa incluem versões do corpus das avaliações efectuadas no contexto do programa ACE (Hachey, Grover e Tobin, 2012), conjuntos de dados construídos a partir da Wikipédia (Culotta, McCallum e Betz, 2006), e subconjuntos de publicações na MEDLINE (Bunescu e Mooney, 2005a). Algumas avaliações conjuntas de sistemas computacionais para o processamento de linguagem natural também abordaram especificamente problemas de extracção de relações semânticas. Um destes eventos foi a tarefa no SemEval 2010 em *Multi-way Classification of Semantic Relations Between Pairs of Nominals* (Hendrickx et al., 2010).

Além dos métodos baseados em aprendizagem supervisionada para a extracção de relações semânticas num dado domínio fechado, importa também referir alguns trabalhos recentes que se focaram na extracção de relações num contexto aberto, tais como o sistema TextRunner[2], o sistema ReVerb[3], o sistema OLLIE[4] ou o SOFIE[5]. Algumas das técnicas elementares que estão na base destes vários sistemas são descritas num artigo de revisão sobre a àrea de *Open Information Extraction* (OIE) (Etzioni et al., 2008).

Os sistemas de OIE procuram identificar um conjunto aberto de relações, operando com base na análise de padrões frequentes em grandes colecções de dados, e muitas vezes usando regras produzidas por peritos humanos. Estes sistemas não requerem assim dados de treino, podendo contemplar a extracção de relações em domínios como a Web (Fader, Soderland e Etzioni, 2011). No entanto, abordagens independentes do domínio produzem geralmente resul-

[2] `http://www.cs.washington.edu/research/textrunner/`
[3] `http://reverb.cs.washington.edu/`
[4] `https://github.com/rbart/ollie`
[5] `http://www.mpi-inf.mpg.de/yago-naga/sofie/`

tados de pior qualidade, e por regra também não normalizam as relações extraídas, tendo este passo de ser tratado à posteriori (Soderland e Mandhani, 2007). Por exemplo, qualquer aplicação usando os resultados destes sistemas deve lidar com homonímia e sinonímia entre as expressões que codificam as relações, assim como com problemas associados à polissemia e sobreposição das relações. Por este motivo, defendemos que deverá ser preferível a extracção de relações semânticas através de exemplos, pelo menos em domínios específicos de aplicação.

Como forma de contornar a dificuldade associada à anotação manual de dados de treino, alguns autores investigaram ainda paradigmas alternativos para a extracção de relações em textos, baseados em supervisão distante, bootstraping (Pantel e Pennacchiotti, 2006) ou outros métodos (Riedel et al., 2013). Por exemplo Mintz et al. (2009), Krause et al. (2012), ou Riedel, Yao e McCallum (2010) usaram a Freebase[6], i.e. uma base de dados estruturada de informação semântica cobrindo milhares de relações entre entidades, como forma de construir os exemplos de treino. Para cada par de entidades referenciado numa relação da Freebase, os autores procuram por frases num corpus que contenham as entidades, usando posteriormente estas frases como exemplos de treino de um extractor de relações baseado num classificador tradicional. Desta forma, os autores combinam as vantagens dos métodos supervisionados para a extracção de relações, com as vantagens dos métodos não-supervisionados para extracção de informação.

Temos, por exemplo, que nas experiências efectuadas por Mintz et al. (2009), os autores utilizaram um classificador de máxima entropia combinando atributos lexicais (e.g., sequências de palavras e as etiquetas morfológicas correspondentes) e sintácticos (e.g., dependências entre as entidades envolvidas na relação). Os resultados mostraram que a metodologia de supervisão distante, com base no Freebase, lhes permitiu extrair 10.000 exemplares de 102 tipos diferentes de relações, com uma precisão de 67.6%. Indo além de recursos como o Freebase, outros autores propuseram ainda a utilização das *infoboxes* da Wikipédia, ou alternativamente de informação proveniente do projecto DBPédia, onde uma rede semântica é extraída automaticamente a partir das *infoboxes* da Wikipédia (Auer et al., 2007), como forma de construir os exemplos de treino (Blessing e Schütze, 2010; Hoffmann, Zhang e Weld, 2010; Wu e Weld, 2010).

[6] `http://www.freebase.com/`

Em relação à extracção de relações de textos em português, temos que alguns resultados foram reportados no contexto da tarefa piloto ReRelEM (Freitas et al., 2008), a qual se realizou no âmbito do segundo evento de avaliação conjunta HAREM (Mota e Santos, 2008) com o objectivo de avaliar o reconhecimento e classificação de relações semânticas ao nível de pares de entidades mencionadas num dado documento. Aquando do evento de avaliação HAREM, a coleção de documentos usada na medição da qualidade dos resultados incluía um total de 12 documentos com 4.417 palavras e 573 entidades mencionadas. Na mesma, encontravam-se descritas 6,790 relações entre pares de entidades manualmente anotadas (i.e., 1.436 relações de identidade, 1.612 relações de ocorrência, 1.232 relações de colocação, e 2.510 relações de outros tipos). Desde então, a colecção de documentos foi estendida e disponibilizada online para outros investigadores (i.e., existem agora 24 tipos de relações semânticas diferentes, e um total de 7.847 entidades mencionadas anotadas na colecção, sendo que 3.776 se encontram relacionadas entre si num total de 4.803 relações manualmente anotadas). No entanto, não temos conhecimento de outros estudos em extracção de relações desde textos, que tenham feito uso desta colecção estendida de documentos e anotações.

Os três sistemas diferentes que participaram na tarefa ReRelEM optaram pelo reconhecimento de diferentes tipos de relações, sendo assim difícil tirar conclusões sobre os seus méritos relativos. Por exemplo o sistema SEI-Geo apenas reconhece relações de inclusão entre entidades do tipo local, realizando esta tarefa através do mapeamento dos pares de entidades para com uma ontologia onde relações semânticas deste tipo já se encontram expressas (Chaves, 2008).

O sistema SeRELeP usa regras heurísticas (e.g., na frase *a Brigada Militar de Porto Alegre ocorre em Porto Alegre*, o sistema iria reconhecer uma relação semântica do tipo ocorrência através de uma regra que indica que se houver uma entidade mencionada do tipo local (e.g., *Porto Alegre*), cujo sintagma seja parte do sintagma de uma entidade mencionada do tipo acontecimento ou organização (e.g., *Brigada Militar de Porto Alegre*), então a entidade inserida é marcada como relacionada com a entidade do tipo local) sobre os resultados do analisador sintáctico PALAVRAS (Bick, 2000), reconhecendo relações de identidade, ocorrência e inclusão entre as entidades mencionadas (Bruckschen et al., 2008).

Finalmente, o sistema REMBRANDT obteve os melhores resultados na tarefa, tendo conseguido uma medida $F_1$ de 45.02%, utilizando

heurísticas básicas de relacionamento entre entidades com base nas suas unidades, nas suas categorias, e nas ligações das respectivas páginas da Wikipédia (e.g., as entidades mencionadas que tenham sido emparelhadas com uma mesma página da Wikipédia são anotadas como sendo idênticas) (Cardoso, 2008).

Além dos esforços realizados no contexto da tarefa ReRelEM, temos que existem também alguns outros trabalhos anteriores que abordaram a tarefa da extracção automática de relações, a partir de textos em português. Por exemplo Oliveira, Costa e Gomes (2010) apresentaram um sistema simples que utiliza padrões léxico-sintácticos para extrair relações de 5 tipos (i.e., *sinonímia*, *hiperonímia*, *parte-de*, *causa* e *finalidade*) entre termos compostos (i.e., termos modificados por adjectivos ou por preposições), a partir de resumos de artigos da versão portuguesa da Wikipédia, desta forma obtendo informação que pode ser utilizada para criar ou estender redes lexicais ao estilo da WordNet.

García e Gamallo (2011) compararam o impacto de diferentes tipos de características linguísticas (i.e., conjuntos de lemas e de etiquetas morfo-sintácticas, padrões léxico-sintácticos, e dependências sintácticas) na tarefa de extracção de relações, através de aprendizagem automática e utilizando uma técnica de supervisão distante que segue o método geral de Mintz et al. (2009) para a recolha de exemplos de treino desde *infoboxes* e textos da Wikipédia. Os autores avaliaram modelos baseados em máquinas de vectores de suporte com diferentes conjuntos de características, através de experiências com dados da versão portuguesa da Wikipédia, e com o foco na extracção de relações do tipo *ocupação* entre pessoas e actividades profissionais.

Resultados preliminares mostraram que as características baseadas nos padrões léxico-sintácticos obtêm uma maior precisão do que as características baseadas em conjuntos de palavras ou dependências sintácticas, muito embora a combinação de diferentes tipos de características ajude a atingir um compromisso entre a precisão e cobertura, melhorando assim sob o desempenho de modelos que apenas utilizem um único tipo de características. As experiências dos autores mostraram também que modelos que usem padrões léxico-sintácticos baseados apenas nos contextos do meio (i.e., que apenas utilizem palavras que ocorram entre os pares de entidades relacionadas) têm um melhor desempenho do que modelos que utilizem todos os contextos (i.e., a informação proveniente de uma janela de palavras ocorrendo antes, entre, e após as entidades mencionadas). Curiosamente, os autores também mencionam que a revisão manual de algumas instâncias (i.e., aquelas que foram posteriormente utilizadas para medir a qualidade das extrações produzidas) revelou que o método de supervisão distante tem uma precisão de cerca de 80% aquando da recolha automática dos exemplos de treino.

Num outro trabalho relacionado, Gamallo, Garcia e Fernández-Lanza (2012) reportam uma abordagem de domínio aberto para a extracção de relações em vários línguas (i.e., os autores abordaram a tarefa de extrair relações entre pares de entidades mencionadas em textos provenientes das versões em inglês, espanhol, português ou galego da Wikipédia), fazendo uso de um analisador sintáctico multilingue baseado em regras. Especificamente, temos que o método de extracção proposto envolve três etapas, nomeadamente (i) a análise de dependências, em que cada frase do texto de entrada é processada com a ferramenta TreeTagger[7] por forma a atribuir etiquetas morfo-sintáticas às palavras, e é de seguida analisada do ponto de vista das dependências com um *parser* multilingue proposto anteriormente e denominado DepPattern[8], (ii) o encontrar das cláusulas constituintes, onde sob cada frase analisada os autores descobrem as cláusulas verbais e os seus constituintes, incluindo as suas funções (e.g., sujeitos, objectos diretos, atributos e complementos preposicionais), e (iii) a aplicação de regras de extracção, onde alguns padrões são usados sobre as cláusulas constituintes para extrair as relações. Temos, por exemplo, que a regra de extracção mais simples é aplicada sobre as cláusulas que contenham apenas um sujeito e um objecto directo. Nestes casos, os dois componentes são os argumentos da relação, enquanto que a expressão verbal corresponde ao tipo da relação. Infelizmente, Gamallo et al. apenas reportam resultados com este método para o caso da língua inglesa, com avaliações iniciais apontando para uma precisão de cerca de 68%.

O trabalho apresentado neste artigo segue a ideia de utilizar supervisão distante como forma de realizar a tarefa de extracção de relações para textos em português, usando especificamente frases da Wikipédia onde co-ocorram entidades relacionadas na DBPédia. Propomos também uma forma diferente de classificar relações semânticas, baseada num método eficiente para a pesquisa por instâncias similares.

[7] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

[8] `http://gramatica.usc.es/pln/tools/deppattern.html`

## 3 Método Proposto

A abordagem proposta para a classificação de uma relação entre duas entidades mencionadas numa frase, de acordo com o seu tipo semântico, é baseada na ideia de encontrar as relações mais semelhantes numa determinada base de dados de relações exemplo previamente anotadas. O procedimento corresponde essencialmente ao desenvolvimento e aplicação de um classificador baseado na votação ponderada dos $kNN$ vizinhos mais próximos, onde cada exemplar de relação tem um peso correspondente à sua semelhança para com a relação a ser classificada. Os exemplares de relações mais similares têm, por consequência, um peso maior na votação, do que os exemplares que são mais dissimilares.

A representação considerada para cada relação exprime-se essencialmente em termos de tetragramas de caracteres, considerando especificamente as palavras que ocorrem:

1. Entre as duas entidades que constituem a relação binária, isto é, entre as duas subsequências correspondentes aos nomes de entidades que são relacionados.
2. Numa janela de três palavras ocorrendo antes da primeira entidade, e entre as entidades envolvidas na relação.
3. Entre as entidades e numa janela de três palavras após a segunda entidade.

Esta representação segue essencialmente a observação de Bunescu e Mooney (2005a) de que uma relação entre duas entidades é geralmente expressa utilizando apenas palavras que aparecem em um de três padrões básicos, nomeadamente antes-e-entre (i.e., palavras antes e entre as duas entidades envolvidas na relação), entre (i.e., apenas as palavras entre as duas entidades), e entre-e-depois (i.e., palavras que ocorrem entre e depois das duas entidades).

Além dos tetragramas de caracteres, também consideramos palavras correspondentes a preposições, verbos e padrões léxico-sintácticos relacionais, que ocorram nas mesmas janelas textuais consideradas para os tetragramas de cáracteres, extraídos com um modelo de etiquitação morfológica desenvolvido com o pacote OpenNLP[9] e treinado com os dados do corpus CINTIL (Branco e Silva, 2006). As preposições e os verbos são extraídos directamente com base nas etiquetas morfológicas. Os padrões relacionais correspondem a uma regra inspirada no sistema de OIE ReVerb (Fader, Soderland e Etzioni, 2011), em que se extraem sequências de palavras formadas por um verbo seguido de uma preposição, ou de um verbo, seguido de vários nomes, adjectivos ou advérbios, e terminando numa preposição.

A cada tetragrama ou palavra, em cada um dos três grupos (isto é, nos grupos antes-e-entre, entre, e entre-e-depois das entidades envolvidas na relação), é atribuído um identificador único. A semelhança entre duas relações pode ser medida através do coeficiente de similaridade de Jaccard entre cada conjunto de identificadores únicos globais, associados às representações.

Muito embora a maioria dos métodos anteriores para extracção de relações usem representações baseadas em palavras individuais, pensamos que a utilização de tetragramas de caracteres pode trazer algumas vantagens, nomeadamente no melhor lidar com problemas de variabilidade lexical. Também experimentámos utilizar outras representações para as relações, utilizando por exemplo $n$-gramas de palavras, depois de lematizar o texto. No entanto, observámos que a representação descrita neste secção consegue o melhor compromisso entre a precisão do classificador e o desempenho computacional.

### 3.1 Geração Automática de Exemplos

A Wikipédia, na sua versão portuguesa para o nosso caso em particular, é um ponto de partida ideal para o desenvolvimento de extractores automáticos de relações, pois trata-se de um recurso abrangente que contém um conjunto muito diversificado de conteúdos aprofundados. Na Wikipédia, além de descrições textuais para conceitos e entidades relevantes, em diferentes domínios do conhecimento, há também informação estruturada sob a forma de *infoboxes*, i.e., tabelas criadas manualmente que apresentam, sob a forma de atributos e valores, factos importantes sobre muitos dos artigos da Wikipédia. Projectos como a DBPédia exploraram a construção automática e a disponibilização de redes de conhecimento derivadas de factos expressos nas *infoboxes* das páginas da Wikipédia em várias línguas, incluindo o português (Auer et al., 2007).

Uma vez que os mesmos factos são frequentemente expressos tanto no texto dos artigos da Wikipédia como nas *infoboxes*, e consequentemente também em recursos como a DBPédia, temos então que combinando as relações da DBPédia com frases constituintes dos artigos na Wikipédia, onde as entidades envolvidas co-ocorram, podemos coleccionar grandes volumes de dados de treino para extractores de relações,

[9] `http://opennlp.apache.org/`

que muito embora sejam ruidosos podem ser úteis dado o seu grande volume (Mintz et al., 2009).

Por exemplo, o artigo da Wikipédia portuguesa sobre o artista *Otis Redding* contém a frase *Otis Redding nasceu na pequena cidade de Dawson, Georgia*. Simultaneamente, a *infobox* deste artigo contém o atributo *origem = Dawson, Georgia* e, consequentemente, a rede da DBPédia contem uma associação do tipo *origem* entre as entidades *Otis Redding* e *Georgia*. Ao combinar a informação da DBPédia com frases dos artigos na Wikipédia, como no exemplo apresentado, podemos gerar dados de treino para um extractor de relações do tipo *origem-de*. Estes dados são muito ruidosos, já que alguns atributos da DBPédia podem não encontrar correspondências em frases da Wikipédia, enquanto outros podem surgir em frases em que as entidades co-ocorrem, mas em que a verdadeira relação não está a ser expressa no texto. No entanto, argumentamos que o grande volume de dados, possível de ser extraído desta forma, compensa o ruído presente nas anotações.

O procedimento geral usado na construção automática da base de dados de exemplos para o extractor de relações é desta forma o seguinte:

1. Recolhem-se da DBPédia todas as relações expressas entre conceitos (i.e., páginas da Wikipédia) correspondentes a pessoas, locais ou organizações. De cada uma destas relações, mantém-se informação sobre as duas entidades que estão relacionadas, e a classe semântica do relacionamento;
2. Para cada relação entre um par de entidades, tal como extraída na primeira etapa, analisamos o texto dos dois artigos da Wikipédia portuguesa correspondentes;
3. O texto dos artigos da Wikipédia é segmentado nas frases constituintes;
4. As frases são filtradas, de modo a manter somente aquelas em que co-ocorrem as duas entidades envolvidas na relação. Este passo de filtragem considera pequenas variações nos nomes das entidades, tal como usados na DBPédia e no nome do artigo da Wikipédia, aquando do mapeamento para com o texto das frases. Desta forma podemos melhorar a abrangência do método proposto. Por exemplo, consideram-se além dos nomes originais, as sequências de caracteres até à primeira vírgula ou parêntesis, dado que muitos conceitos da Wikipédia são desambiguados através da inclusão de mais informação no nome – por exemplo, a página da Wikipédia correspondente ao estado da Georgia nos EUA, é identificada pelo sequência de caracteres *Georgia_(Estados_Unidos)*, embora seja de esperar que muitas frases apenas se refiram a este estado pelo nome de *Georgia*.
5. As frases que resultam da etapa de filtragem anterior são mantidas como exemplares de um determinado tipo de relação semântica.

Após a execução do procedimento descrito acima, vamos obter muito exemplos dos vários tipos de relações semânticas que se encontram codificados na DBPédia, os quais foram por sua vez derivados da informação nas *infoboxes* da Wikipédia. Uma vez que muitos destes tipos de relações correspondem a ligeiras variações de um mesmo conceito semântico (e.g., *locatedInArea* ou *subRegion* são variações de um mesmo conceito que se pode generalizar para *localizado-em*), procedemos a um agrupamento/generalização manual dos diferentes tipos de relações presentes na DBPédia, tendo finalmente obtido um conjunto de dados contendo 10 tipos de relações diferentes, tal como ilustrado na Tabela 1. Importa referir que as associações entre os oito primeiros tipos de relações na Tabela 1 são orientadas (i.e., estes tipos de relações devem ser consideradas como assimétricas), enquanto que as associações entre os últimos dois tipos (i.e., relações do tipo *parceiro* e *não-relacionado/outros*) são simétricas.

Tal como referido atrás, temos que um pequeno sub-conjunto dos exemplos de treino gerados automaticamente foi posteriormente revisto manualmente, por forma a construir uma colecção para a avaliação de resultados. Durante este processo de revisão manual, e ainda que de uma forma muito informal, verificou-se que o método de supervisão distante tem uma precisão de cerca de 80% na atribuição de tipos de relações que se encontrem realmente expressos nas frases, aquando da recolha automática dos exemplos de treino. Este resultado está em concordância com o trabalho anterior de García e Gamallo (2011). Importa no entanto referir que, também aquando

| **Relação** | **Núm. Exemplos** |
|---|---|
| local-de-enterro-ou-falecimento | 6.726 |
| influenciado-por | 147 |
| pessoa-chave-em | 355 |
| localizado-em | 46.236 |
| origem-de | 23.664 |
| antepassado-de | 266 |
| parte-de | 5.142 |
| sucessor-de | 496 |
| parceiro | 128 |
| não-relacionado/outros | 6.441 |

Tabela 1: Tipos de relações considerados.

| Tipo de Relação | Exemplos de Instâncias de Relações |
|---|---|
| local-de-enterro-ou-falecimento | **Camilo Pessanha** morreu no dia 1 de Março de 1926 em **Macau**, devido ao uso excessivo de Ópio.<br>Classe DBPédia : **deathPlace** Direcção : (entidade1,entidade2) |
| | **Corisco** foi enterrado em Jeremoabo, na **Bahia**.<br>Classe DBPédia : **placeOfBurial** Direcção : (entidade2,entidade1) |
| influênciado-por | O som inicial do **U2** foi influenciado por bandas como **Television** e Joy Division.<br>Classe DBPédia : **influencedBy** Direcção : (entidade1,entidade2) |
| | **Rubem Fonseca** escreveu os contos "Chegou o Outono", "Noturno de Bordo"e "Mistura" baseado na linguagem de **Machado de Assis**.<br>Classe DBPédia : **influenced** Direcção : (entidade2,entidade1) |
| pessoa-chave-em | **Magic Circle Music** foi fundada pelo baixista do Manowar **Joey DeMaio** em 2005.<br>Classe DBPédia : **foundedBy** Direcção : (entidade1,entidade2) |
| | A **Microsoft** foi fundada em 1975 por **Bill Gates** e Paul Allen.<br>Classe DBPédia : **keyPerson** Direcção : (entidade2,entidade1) |

Tabela 2: Exemplos para alguns dos diferentes tipos de relações consideradas.

| Tipo de Relação | Padrões Relacionais Extraídos |
|---|---|
| origem | nasceu em; começou a; competiu em; nasceu a; foi formado em; |
| influenciado-por | é inspirada por; combinando com; apareceu em; influenciou ministérios de; foi influenciado por; |
| local-de-enterro-ou-falecimento | nasceu em; morreu em; faleceu em; visconde com; morreu de; |
| parceiro | é casado com; foi casado com; casou com; casou-se com; compete ao lado; |

Tabela 3: Os 5 padrões relacionais mais frequentes para alguns dos tipos de relações semânticas.

do processo de revisão manual, foram detectados vários problemas ao nível da segmentação das frases provenientes dos artigos da Wikipédia (e.g., é comum observar frases que incluem, no seu início ou no final, palavras provenientes do título da secção imediatamente antes da frase). Na construção da colecção manualmente revista para a avaliação de resultados, todos os problemas detectados foram corrigidos.

A Tabela 2 mostra alguns exemplares das diferentes classes de relações consideradas após a generalização, mostrando ainda a classe da relação originalmente expressa na DBPédia, assim como a direcção do relacionamento.

Por outro lado a Tabela 3 mostra alguns dos padrões relacionais mais frequentemente associados a algumas das relações semânticas consideradas, dando assim uma ideia dos valores das características léxico-sintácticas que foram usadas nas representações dos exemplares de relações.

### 3.2 Pesquisa por Relações Similares

Importa observar que uma abordagem simplista para encontrar os exemplares de relações mais similares entre si, numa base de dados de tamanho $N$, envolve o cálculo da similaridade entre $N^2$ pares de exemplares. Este procedimento torna-se rapidamente difícil de escalar para valores grandes de $N$. Apesar de a tarefa ser paralelizável, é necessário baixar a complexidade $O(N^2)$ para alcançar uma boa escalabilidade. Desta forma, o desenho de operações de pré-processamento adequadas, que facilitem os cálculos de similaridade entre exemplares, assume uma importância relevante. No nosso método, isto é feito pelo cálculo de uma aproximação ao coeficiente de similaridade de Jaccard, obtida através de uma técnica baseada em valores mínimos de funções de dispersão (i.e., *min-hash*), e utilizando ainda uma técnica de dispersão sensível à localização (*Locality-Sensitive Hashing (LSH)*) para encontrar rapidamente as $kNN$ relações mais similares.

A técnica de *min-hash* foi apresentada no trabalho seminal de Broder (1997; Broder et al. (2000)), onde os autores descrevem uma aplicação bem sucedida na detecção de páginas Web duplicadas. Dado um vocabulário $\Omega$ de tamanho $D$ (ou seja, o conjunto de todos os elementos representativos usados nas descrições dos exemplares de relações), e dois conjuntos de elementos, $S_1$ e $S_2$, onde $S_1, S_2 \subseteq \Omega = \{1, 2, \ldots, D\}$

temos que o coeficiente de similaridade de Jaccard, entre os dois conjuntos de elementos, é dado pela razão entre o tamanho da intersecção de $S_1$ e $S_2$, sobre o tamanho da sua união:

$$\begin{aligned} \mathrm{J}(S_1, S_2) &= \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \\ &= \frac{|S_1 \cap S_2|}{|S_1| + |S_2| - |S1 \cap S_2|} \end{aligned} \quad (1)$$

Os dois conjuntos são mais semelhantes entre si quando o coeficiente de similaridade de Jaccard está perto de 1, e mais dissimilares quando o coeficiente de Jaccard é próximo de 0.

Para pares de conjuntos grandes, calcular eficientemente os tamanhos dos conjuntos resultantes da sua união e intersecção é computacionalmente exigente, uma vez que o número total de elementos a analisar é enorme. No entanto, suponhamos que uma permutação aleatória $\pi$ é realizada sobre o vocabulário $\Omega$, ou seja:

$$\pi : \Omega \longrightarrow \Omega \ , onde \ \Omega = \{1, 2, \ldots, D\}. \quad (2)$$

Notando que o coeficiente de Jaccard corresponde à razão entre o número de elementos que ocorre simultaneamente em $S_1$ e $S_2$, sobre o número de elementos que ocorre em pelo menos um dos conjuntos, temos que um argumento elementar de análise de probabilidades pode mostrar que:

$$\Pr\left(\min(\pi(S_1)) = \min(\pi(S_2))\right) = \mathrm{J}(S_1, S_2) \quad (3)$$

Após a criação de $k$ permutações independentes dos elementos pertencentes a $\Omega$, pode-se estimar a medida de similaridade $\mathrm{J}(S1, S2)$ de forma eficiente e não tendenciosa, como uma distribuição amostral de uma variável aleatória binomial:

$$\begin{aligned} \hat{\mathrm{J}}(S_1, S_2) &= \frac{1}{k} \sum_{j=1}^{k} \mathrm{um}(\min(\pi_k(S_1)) \\ &= \min(\pi_k(S_2))) \end{aligned} \quad (4)$$

Na fórumula acima a função um() devolve o valor de 1 quando para uma dada permutação o elemento mínimo dos dois conjuntos é igual, e o valor de 0 caso contrário.

$$\mathrm{Var}(\hat{\mathrm{J}}(S_1, S_2)) = \frac{1}{k} \mathrm{J}(S_1, S_2)\left(1 - \mathrm{J}(S_1, S_2)\right) \quad (5)$$

Na implementação do mecanismo de *min-hash*, cada uma das permutações independentes corresponde a um valor de uma função de dispersão, no nosso caso considerando 32 *bits* de armazenamento. Cada uma das permutações independentes $k$ está assim associada a uma função de *hash* polinomial $h^k(x)$ que mapeia os membros de $\Omega$ para valores distintos. Para qualquer conjunto $S$, tomarmos os $k$ valores de $\mathrm{h}^k_{min}(S)$, ou seja, o membro de $S$ com o valor mínimo de $h^k(x)$. O conjunto de $k$ valores é referido como a assinatura *min-hash* de um exemplo.

A pesquisa eficiente pelos $kNN$ vizinhos mais próximos é implementada através de uma técnica simples de dispersão sensível a localização, que utiliza as assinaturas de *min-hash* para comprimir as representações das relações em assinaturas pequenas (ou seja, para gerar assinaturas pequenas, do conjunto de todos os tetragramas de caracteres, preposições, verbos, e padrões relacionais, ocorrendo antes-e-entre, entre, e entre-e-depois das entidades envolvidas na relação), ao mesmo tempo preservando a similaridade esperada de qualquer par de instâncias. Esta técnica utiliza $L$ tabelas de dispersão diferentes (ou seja, na nossa implementação, usamos $L$ estruturas de dados persistentes construídas com a biblioteca MapDB[10]), cada uma correspondendo a um $n$-tuplo das assinaturas *min-hash*, a que nos referimos aqui como uma banda. No momento da classificação, calculamos a assinatura *min-hash* da relação a ser classificada, e de seguida consideramos qualquer relação de exemplo que se encontre associada a um mesmo contentor da estrutura de dados, para qualquer uma das bandas *min-hash*, como uma relação candidata a pertencer ao conjunto das $kNN$ mais similares. Verificamos apenas os pares candidatos, utilizando as assinaturas *min-hash* completas para aproximar o coeficiente de similaridade de Jaccard. Desta forma, podemos evitar as comparações de similaridade com todas as relações na base de dados de exemplos. O Capítulo 3 do livro de Rajaraman e Ullman (2011) descreve o uso da assinaturas *min-hash* com técnicas baseadas em dispersão sensível a localização, em aplicações relacionadas com a pesquisa por itens semelhantes.

Um esboço completo do método de classificação proposto é assim o seguinte. Começando por analisar o conjunto de frases envolvido na indexação dos exemplares de treino:

1. Extraem-se conjuntos de tetragramas de caracteres, preposições, verbos, e padrões relacionais das *substrings* que ocorrem antes-e-entre, entre, e entre-e-depois das entidades envolvidas na relação, para cada relação em cada frase de um determinado conjunto de textos de exemplo. Os exemplos são previamente recolhidos da Wikipédia, tal como explicado na subsecção anterior, correspondendo a frases onde co-ocorrem pares de en-

[10] `http://www.mapdb.org/`

tidades relacionados na DBPédia.

2. As assinaturas *min-hash* são extraídas a partir dos conjuntos gerados na primeira etapa.

3. As assinaturas são divididas em bandas, e os exemplares de relações que estas representam são indexados em $L$ diferentes tabelas de dispersão, com base nos valores presentes nas bandas das assinaturas.

Na classificação de relações, e para verificar se uma dada relação semântica está ou não descrita numa frase, seguem-se os seguintes passos:

1. Começamos pela extracção dos tetragramas de caracteres, preposições, verbos, e padrões de relacionamento, a partir das subsequências que ocorrem antes-e-entre, entre, e entre-e-depois das entidades envolvidas.

2. Gera-se uma assinatura *min-hash* a partir do conjunto gerado no primeiro passo.

3. As relações de exemplo com pelo menos uma banda idêntica no índice construído na fase de indexação são consideradas como candidatas, e a sua semelhança para com a relação a classificar é, então, estimada usando as assinaturas *min-hash* completas.

4. Os exemplos mais semelhantes são mantidos numa lista de prioridades, de onde posteriormente se podem extrair os $kNN$ exemplares mais semelhantes.

5. Os $kNN$ exemplares mais semelhantes são analisados, e a classe semântica da relação é atribuída com base numa votação, ponderada pelo valor de similaridade, entre as classes presentes nos $kNN$ exemplares mais semelhantes.

## 4 Avaliação Experimental

O método de extracção de relações aqui proposto foi avaliado com base em frases da Wikipédia e de duas formas distintas, nomeadamente:

1. Deixando de fora da fase de indexação, na base de dados de exemplos, as relações correspondentes a uma pequena parte dos dados gerados automaticamente a partir da Wikipédia, correspondente a um conjunto de relações de exemplo verificadas manualmente quanto à sua exactidão.

2. Deixando de fora da fase de indexação 25% dos exemplos de cada classe semântica.

| | **Conjuntos de Dados** | | |
|---|---|---|---|
| | Treino | Teste | Total |
| # Frases | 97.363 | 625 | 97.988 |
| # Palavras | 2.172.125 | 14.320 | 2.186.445 |
| # Classes | 10 | 10 | 10 |
| # Instâncias | 89.054 | 547 | 89.601 |
| # Entidades únicas | 70.716 | 838 | 71.119 |
| Média palavras/frase | 22,42 | 24,12 | 22.43 |
| StDev. palavras/frase | 11.39 | 11.00 | 11.39 |
| Média instâncias/classe | 8.905,4 | 54,7 | 8.960,1 |
| StDev. instâncias/classe | 14,109.33 | 64.18 | 14.172,38 |

Tabela 4: Caracterização estatística dos conjuntos de dados usados nas diferentes experiências.

A Tabela 4 apresenta uma caracterização estatística do sub-conjunto dos dados que se encontra manualmente verificado (i.e., a coluna assinalada como teste), assim como do sub-conjunto dos dados para os quais não temos anotações manuais (i.e., a coluna assinalada como treino, correspondendo às relações de exemplo que são indexadas nos testes relacionados com o primeiro método experimental), e para o conjunto completo de exemplares de relações. A anotação manual consistiu em verificar se, de facto, as frases que estavam a ser geradas através do processo automatizado correspondem verdadeiramente a exemplos válidos de um tipo semântico de uma relação em particular. O conjunto de dados completo usado nas nossos experiências encontra-se disponibilizado online[11].

Realizámos experiências com representações diferentes das relações (por exemplo, utilizando apenas tetragramas de caracteres, ou usando tetragramas e as características derivadas de etiquetas morfológicas), e também com diferentes parâmetros no método de classificação baseado em assinaturas *min-hash*, através da variação do número de vizinhos mais próximos que foi considerado (i.e., 1, 3, 5 ou 7), variando o tamanho das assinaturas *min-hash* (i.e., 200, 400, 600 ou 800 números inteiros) e o número de bandas LSH considerado (i.e., 25 ou 50 bandas). Importa notar que, ao usar $b$ bandas LSH, cada uma com $r$ valores, temos que a probabilidade de as assinaturas *min-hash* de dois conjuntos $S_1$ e $S_2$ concordarem em todas os valores de pelo menos uma banda, gerando-se assim um par candidato, é de $1-(1-\mathrm{J}(S_1,S_2)^r)^b$. Com 50 bandas e uma assinatura *min-hash* de tamanho 600, cerca de um em cada mil pares com uma similaridade até 85% vai deixar de se tornar um par candidato através do método LSH e, por consequência, vai ser um falso negativo. Especificamente com estes parâmetros, as relações com uma similaridade abaixo de 85% são muito susceptíveis de ser

[11] `http://dmir.inesc-id.pt/project/DBpediaRelations-PT_01_in_English`

descartadas através do método LSH, o que pode contribuir para a confiança em uma classificação correta (ou seja, estamos de certa forma a trocar precisão por abrangência, nos parâmetros considerados para a indexação).

Como medidas de avaliação, utilizámos principalmente as macro-médias da precisão (P), abrangência (A), e da medida $F_1$ sobre todos os tipos de relações, excepto o tipo *não-relacionado/outro*. Usamos assim a macro-média das pontuações sob 18 classes de relações semânticas, uma vez que temos duas direcções possíveis para 8 tipos semânticos das relações inferidas a partir DBPédia (i.e., as relações *parceiro* e *não-relacionado/outro* são bidireccionais).

A Tabela 5 apresenta os resultados obtidos para diferentes representações e parâmetros de indexação, quando se considera o conjunto de dados com as anotações manuais. A Tabela 6 apresenta os resultados obtidos para diferentes parâmetros de indexação com o conjunto completo de características, sob 25% do conjunto completo de relações de cada classe. Os resultados mostram que o método usando supervisão distante, juntamente com a técnica de classificação proposta, permite extrair relações com uma exactidão razoável. Também podemos verificar que os valores das diferentes métricas de avaliação são ligeiramente inferiores no caso dos testes com os 25% do conjunto total de exemplos. Isto indica que os resultados medidos com a colecção manualmente anotada podem ser encarados como um limite superior a uma aproximação da verdadeira exactidão do sistema.

Os resultados da Tabela 5 indicam também que a combinação de tetragramas de caracteres, verbos, preposições, e padrões relacionais, proporciona um melhor desempenho de identificação e classificação. Os resultados sugerem ainda que a utilização dos cinco ou sete primeiros vizinhos, em vez de apenas o exemplo mais semelhante, resulta num aumento de desempenho.

A Tabela 7 apresenta resultados individuais por classe sob 25% dos exemplares de cada relação, considerando as características de representação e indexação que obtiveram o melhor desempenho nos resultados das Tabelas 5 e 6. Isto corresponde a uma configuração com:

- Tetragramas de caracteres, verbos, preposições e padrões relacionais para representar as relações semânticas;
- Assinaturas *min-hash* com tamanho = 800;
- Número de bandas no método LSH = 25;
- Os sete vizinhos mais próximos;

Além dos resultados para a configuração regular de classificação de relações de acordo com os tipos e com a direcção, também apresentamos resultados para uma avaliação em que se ignoram as direcções das relações, bem como os resultados obtidos para a classe correspondente ao tipo *não-relacionado/outro*. Finalmente, esta tabela apresenta também uma avaliação global dos resultados obtidos através da medida de exactidão, a qual mede a porção de classificações corretas, dando assim uma maior importância aos tipos de relações com maior número de ocorrências no corpus, o que não acontece com as macro-médias. Os resultados mostram que classes como *origem-de* e *parte-de* são relativamente fáceis de identificar e classificar, enquanto que classes como *influenciado-por* ou *sucessor-de* são muito mais difíceis de identificar e classificar correctamente. Note-se por exemplo que, para a classe correspondente a *influenciado-por*, o conjunto de dados indexado contém apenas 110 relações de exemplo, enquanto que o conjunto de exemplos usado para a medição de resultados nesta classe tem apenas 35 exemplares de relações.

## 5 Conclusões e Trabalho Futuro

A utilização de técnicas de Extracção de Informação como forma de suportar a criação de bases de conhecimento, em larga escala, a partir de repositórios de documentos de texto, tais como a Web ou como colecções de textos jornalísticos, é objecto actual de estudo intenso. No entanto, as melhores abordagens existentes, para a extracção de relações semânticas, não são facilmente transponíveis para línguas ou domínios diferentes. Temos ainda que os métodos supervisionados requerem grandes quantidades de dados anotados, e têm uma complexidade computacional elevada. Por outro lado, as técnicas independentes de domínio apresentam resultados de baixa precisão e não normalizam as relações.

Neste artigo foi proposta uma abordagem de supervisão distante para a classificação de relações extraídas de textos escritos em português, suportada por dados extraídos da Wikipédia e da DBPédia, e baseada na medição de similaridade entre as relações a classificar e relações armazenadas numa base de dados de relações de exemplo. No método proposto, os exemplos de treino são recolhidos automaticamente a partir da Wikipédia, correspondendo a frases que expressam relações entre pares de entidades extraídas da DBPédia. Estes exemplos são representados como assinaturas *min-hash* de conjuntos de elementos, originalmente contendo

| Características | Min Hash | 1 kNN P | 1 kNN A | 1 kNN $F_1$ | 3 kNN P | 3 kNN A | 3 kNN $F_1$ | 5 kNN P | 5 kNN A | 5 kNN $F_1$ | 7 kNN P | 7 kNN A | 7 kNN $F_1$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tetragramas | 200/25 | 0.492 | 0.400 | 0.441 | 0.627 | 0.426 | 0.507 | 0.716 | 0.423 | 0.532 | 0.724 | 0.429 | 0.539 |
| | 200/50 | 0.489 | 0.400 | 0.440 | 0.625 | 0.425 | 0.506 | 0.716 | 0.423 | 0.532 | 0.726 | 0.430 | 0.540 |
| | 400/25 | 0.476 | 0.405 | 0.438 | 0.559 | 0.418 | 0.478 | 0.724 | 0.434 | 0.543 | **0.736** | **0.443** | **0.553** |
| | 400/50 | 0.474 | 0.405 | 0.437 | 0.557 | 0.423 | 0.481 | 0.715 | 0.434 | 0.540 | 0.731 | 0.441 | 0.550 |
| | 600/25 | 0.609 | 0.435 | 0.508 | 0.645 | 0.437 | 0.521 | 0.688 | 0.440 | 0.537 | 0.663 | 0.440 | 0.529 |
| | 600/50 | 0.583 | 0.435 | 0.498 | 0.646 | 0.437 | 0.521 | 0.686 | 0.433 | 0.531 | 0.719 | 0.441 | 0.547 |
| | 800/25 | 0.545 | 0.426 | 0.478 | 0.610 | 0.430 | 0.504 | 0.651 | 0.434 | 0.521 | 0.640 | 0.442 | 0.523 |
| | 800/50 | 0.541 | 0.423 | 0.475 | 0.611 | 0.432 | 0.506 | 0.652 | 0.436 | 0.523 | 0.643 | 0.444 | 0.525 |
| Tetragramas e Verbos | 200/25 | 0.476 | 0.414 | 0.443 | 0.628 | 0.437 | 0.515 | 0.713 | 0.429 | 0.536 | 0.718 | 0.432 | 0.539 |
| | 200/50 | 0.474 | 0.414 | 0.442 | 0.628 | 0.437 | 0.515 | 0.713 | 0.429 | 0.536 | 0.718 | 0.432 | 0.539 |
| | 400/25 | 0.499 | 0.417 | 0.454 | 0.563 | 0.430 | 0.488 | 0.725 | 0.437 | 0.545 | 0.729 | 0.442 | 0.550 |
| | 400/50 | 0.497 | 0.417 | 0.453 | 0.565 | 0.436 | 0.492 | 0.674 | 0.440 | 0.532 | 0.729 | 0.443 | 0.551 |
| | 600/25 | 0.580 | 0.425 | 0.491 | 0.640 | 0.442 | 0.523 | 0.669 | 0.439 | 0.530 | 0.728 | 0.435 | 0.545 |
| | 600/50 | 0.553 | 0.425 | 0.481 | 0.641 | 0.442 | 0.523 | 0.724 | 0.439 | 0.547 | 0.728 | 0.441 | 0.549 |
| | 800/25 | 0.549 | 0.424 | 0.479 | 0.615 | 0.433 | 0.508 | 0.720 | 0.443 | 0.549 | **0.736** | 0.441 | **0.551** |
| | 800/50 | 0.549 | 0.424 | 0.479 | 0.615 | 0.433 | 0.508 | 0.712 | **0.447** | 0.549 | 0.731 | 0.438 | 0.548 |
| Tetragramas, Verbos e Preposições | 200/25 | 0.477 | 0.403 | 0.437 | 0.628 | 0.431 | 0.511 | 0.720 | 0.432 | 0.540 | 0.723 | 0.438 | 0.546 |
| | 200/50 | 0.478 | 0.404 | 0.438 | 0.628 | 0.431 | 0.511 | 0.666 | 0.432 | 0.524 | 0.670 | 0.438 | 0.530 |
| | 400/25 | 0.522 | 0.431 | 0.472 | 0.574 | 0.432 | 0.493 | 0.732 | 0.446 | 0.554 | 0.731 | 0.442 | 0.551 |
| | 400/50 | 0.522 | 0.431 | 0.472 | 0.578 | 0.441 | 0.500 | 0.679 | 0.446 | 0.538 | 0.732 | 0.445 | 0.554 |
| | 600/25 | 0.581 | 0.427 | 0.492 | 0.630 | 0.432 | 0.513 | 0.673 | 0.446 | 0.536 | 0.677 | 0.441 | 0.534 |
| | 600/50 | 0.554 | 0.427 | 0.482 | 0.631 | 0.432 | 0.513 | 0.726 | 0.439 | 0.547 | 0.731 | 0.442 | 0.551 |
| | 800/25 | 0.548 | 0.426 | 0.479 | 0.616 | 0.435 | 0.510 | 0.721 | **0.449** | 0.553 | **0.733** | 0.447 | **0.555** |
| | 800/50 | 0.545 | 0.423 | 0.476 | 0.620 | 0.446 | 0.519 | 0.721 | 0.445 | 0.550 | 0.732 | 0.446 | 0.554 |
| Tetragramas, Verbos, Preposições e Padrões Relacionais | 200/25 | 0.472 | 0.404 | 0.435 | 0.629 | 0.436 | 0.515 | 0.724 | 0.436 | 0.544 | 0.723 | 0.440 | 0.547 |
| | 200/50 | 0.474 | 0.404 | 0.436 | 0.575 | 0.436 | 0.496 | 0.671 | 0.436 | 0.529 | 0.670 | 0.440 | 0.531 |
| | 400/25 | 0.521 | 0.429 | 0.471 | 0.572 | 0.429 | 0.490 | 0.730 | 0.443 | 0.551 | 0.731 | 0.441 | 0.550 |
| | 400/50 | 0.521 | 0.429 | 0.471 | 0.573 | 0.436 | 0.495 | 0.680 | 0.447 | 0.539 | **0.732** | 0.444 | 0.553 |
| | 600/25 | 0.579 | 0.423 | 0.489 | 0.628 | 0.429 | 0.510 | 0.673 | 0.446 | 0.536 | 0.678 | 0.437 | 0.531 |
| | 600/50 | 0.552 | 0.423 | 0.479 | 0.629 | 0.428 | 0.509 | 0.728 | 0.446 | 0.553 | 0.731 | 0.438 | 0.548 |
| | 800/25 | 0.547 | 0.423 | 0.477 | 0.616 | 0.433 | 0.509 | 0.715 | 0.445 | 0.549 | 0.723 | 0.444 | 0.550 |
| | 800/50 | 0.544 | 0.420 | 0.474 | 0.618 | 0.439 | 0.513 | 0.716 | 0.444 | 0.548 | 0.731 | **0.449** | **0.556** |

Tabela 5: Resultados para diferentes representações das relações e parâmetros de indexação.

| Características | Min Hash | 1 kNN P | 1 kNN A | 1 kNN $F_1$ | 3 kNN P | 3 kNN A | 3 kNN $F_1$ | 5 kNN P | 5 kNN A | 5 kNN $F_1$ | 7 kNN P | 7 kNN A | 7 kNN $F_1$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tetragramas, Verbos, Preposições e Padrões Relacionais | 200/25 | 0.448 | 0.353 | 0.395 | 0.460 | 0.345 | 0.394 | 0.492 | 0.331 | 0.396 | 0.487 | 0.325 | 0.390 |
| | 200/50 | 0.450 | 0.354 | 0.396 | 0.459 | 0.347 | 0.395 | 0.489 | 0.332 | 0.395 | 0.507 | 0.328 | 0.398 |
| | 400/25 | 0.440 | 0.350 | 0.390 | 0.448 | 0.344 | 0.389 | 0.468 | 0.328 | 0.386 | 0.479 | 0.320 | 0.384 |
| | 400/50 | 0.439 | 0.351 | 0.390 | 0.445 | 0.343 | 0.387 | 0.465 | 0.327 | 0.384 | 0.483 | 0.321 | 0.386 |
| | 600/25 | 0.461 | 0.358 | 0.403 | 0.466 | 0.353 | 0.401 | 0.482 | 0.337 | 0.397 | 0.469 | 0.324 | 0.383 |
| | 600/50 | 0.461 | **0.360** | 0.404 | 0.463 | 0.353 | 0.401 | 0.490 | 0.340 | 0.401 | 0.492 | 0.329 | 0.394 |
| | 800/25 | 0.446 | 0.358 | 0.397 | 0.462 | 0.350 | 0.398 | 0.492 | 0.338 | 0.401 | **0.516** | 0.333 | **0.405** |
| | 800/50 | 0.445 | 0.358 | 0.397 | 0.453 | 0.349 | 0.394 | 0.484 | 0.336 | 0.397 | 0.510 | 0.333 | 0.403 |

Tabela 6: Resultados obtidos sob 25% do conjunto completo de instâncias de cada classe.

tetragramas de caracteres assim como outros elementos representativos, e indexados numa estrutura de dados que implementa a ideia de dispersão sensível a localização. Para verificar qual a relação semântica que se encontra expressa entre um determinado par de entidades, são procurados os $kNN$ exemplos de treino mais similares, e a relação é atribuída com base numa votação ponderada. Testes com um conjunto de dados da Wikipédia comprovam a adequabilidade do método proposto, sendo que o mesmo é, por exemplo, capaz de extrair 10 tipos diferentes de relações semânticas, oito deles correspondendo a tipos de relações assimétricos, com uma pontuação média de 55.6% em termos da medida $F_1$.

Apesar dos resultados interessantes, há também muitos desafios em aberto para trabalho futuro. Temos, por exemplo, que a maioria dos métodos baseados em *kernels*, do actual estado-da-arte, exploram semelhanças entre representações de relações baseadas em grafos, derivados simultâneamente de informações lexicais e de estruturas resultantes de uma análise sintáctica e de dependências (Nguyen, Moschitti

| **Relação** | **Direcção** | **Instâncias** (treino/teste) | **Assimétricas** P | A | $F_1$ | **Simétricas** P | A | $F_1$ |
|---|---|---|---|---|---|---|---|---|
| local-de-enterro-ou-falecimento | (e1,e2) | 4.788/1.596 | 0.802 | 0.595 | 0.683 | 0.806 | 0.574 | 0.671 |
| | (e2,e1) | 257/85 | 0.375 | 0.035 | 0.065 | | | |
| influenciado-por | (e1,e2) | 84/28 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | (e2,e1) | 26/9 | 1.000 | 0.111 | 0.199 | | | |
| pessoa-chave-em | (e1,e2) | 106/35 | 0.500 | 0.086 | 0.146 | 0.233 | 0.079 | 0.117 |
| | (e2,e1) | 161/53 | 0.200 | 0.113 | 0.145 | | | |
| localizado-em | (e1,e2) | 33.639/11.213 | 0.916 | 0.929 | 0.922 | 0.924 | 0.922 | 0.923 |
| | (e2,e1) | 1.038/346 | 0.395 | 0.087 | 0.142 | | | |
| origem-de | (e1,e2) | 16.784/5.594 | 0.723 | 0.806 | 0.807 | 0.733 | 0.908 | 0.811 |
| | (e2,e1) | 965/321 | 0.664 | 0.567 | 0.612 | | | |
| antepassado-de | (e1,e2) | 151/50 | 0.471 | 0.800 | 0.593 | 0.545 | 0.727 | 0.623 |
| | (e2,e1) | 49/16 | 0.000 | 0.000 | 0.000 | | | |
| parte-de | (e1,e2) | 2.590/863 | 0.541 | 0.544 | 0.543 | 0.680 | 0.576 | 0.623 |
| | (e2,e1) | 1.267/422 | 0.574 | 0.275 | 0.372 | | | |
| sucessor-de | (e1,e2) | 117/39 | 0.400 | 0.051 | 0.091 | 0.541 | 0.161 | 0.248 |
| | (e2,e1) | 255/85 | 0.359 | 0.165 | 0.226 | | | |
| parceiro | — | 96/32 | — | — | — | 0.600 | 0.188 | 0.286 |
| não-relacionado/outros | — | 4.831/1.610 | — | — | — | 0.767 | 0.543 | 0.636 |
| Macro-médias | — | — | 0.516 | 0.333 | 0.405 | 0.583 | 0.468 | 0.494 |
| Exactidão | — | — | | 0.813 | | | 0.834 | |

Tabela 7: Resultados obtidos individualmente para cada classe e direcção de relacionamento, sob 25% do conjunto completo de instâncias de cada classe.

e Riccardi, 2009). Estudos recentes têm proposto métodos baseados em assinaturas *min-hash* para comparar grafos (Teixeira, Silva e Jr., 2012). Para trabalho futuro, seria interessante experimentar a aplicação destes métodos na tarefa de extracção de relações em textos, usando desta forma representações ricas para os exemplos de relações, baseadas em grafos.

Desde o trabalho seminal de Broder (1997) sobre a utilização de assinaturas *min-hash* para a detecção de páginas Web duplicadas, ocorreram desenvolvimentos teóricos e metodológicos consideráveis, em termos da aplicação deste tipo de abordagens. Para trabalho futuro, gostaríamos de avaliar a abordagem $b$-bit *minwise hashing* de Li e König (2010) para melhorar a eficiência de armazenamento, experimentar com a extensão proposta por Chum, Philbin e Zisserman (2008) para aproximar medidas de similaridade entre histogramas de valores, e com uma abordagem em duas etapas semelhante à do sistema de desambiguação de entidades KORE (Hoffart et al., 2012), onde documentos textuais são representados por frases chave, que por sua vez são representados como conjuntos de $n$-gramas.

Finalmente gostaríamos de realizar experiências, com o método proposto neste artigo, sobre outras colecções de dados, de forma a avaliar a técnica de extracção de relações em textos de outros géneros, tais como artigos técnicos, textos literários, documentos jurídicos, etc. Em particular, seria interessante aplicar o método proposto à colecção de textos do ReRelEM, como forma de validar o método de supervisão distante proposto neste artigo. Gostaríamos assim de experimentar com dados derivados de um mapeamento das classes da DBPédia para com as classes do ReRelEM, permitindo assim a validação dos resultados.

Ainda no que se refere a experiências com outros tipos de dados, importa referir que embora apenas tenhamos feito algumas experiências iniciais relacionadas com a utilização do método proposto na extracção de relações em textos provenientes de outros domínios (e.g., usando frases da Wikipédia como dados de treino e tentando extrair relações em frases provenientes de textos jornalísticos, posteriormente observando a qualidade dos resultados de um modo informal), importa referir que os textos da Wikipédia constituem um género muito específico, onde existem determinados padrões que são frequentemente usados como forma de expressar relações (e.g., as relações do tipo *origem-de* são tipicamente expressas à custa de um padrão em que a primeira menção a um determinado nome de pessoa é seguida do local e data de nascimento, entre parêntesis). Em textos de outros domínios, o método proposto vai muito provavelmente obter resultados diferentes em termos da qualidade

das extracções, sendo que os nossos testes iniciais apontam no sentido de ser difícil vir a usar frases da Wikipédia como forma de aprender bons extractores de relações para outros domínios.

## Agradecimentos

Este trabalho foi suportado pela Fundação para a Ciência e Tecnologia (FCT), através do projecto com referência PTDC/EIA-EIA/109840/2009 (SInteliGIS), assim como através dos projectos com referências PTDC/EIA-EIA/115346/200912 (SMARTIES), UTA-Est/MAI/0006/2009 (REACTION), e através do financiamento plurianual do laboratório associado INESC-ID com a referência PEst-OE/EEI/LA0021/2013. O autor David Batista foi também suportado pela bolsa de doutoramento da FCT com referência SFRH/BD/70478/2010.

## Referências

Airola, Antti, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, e Tapio Salakoski. 2008. A graph kernel for protein-protein interaction extraction. Em *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*.

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, e Zachary Ives. 2007. DBpedia: a nucleus for a web of open data. Em *Proceedings of the International Conference on the Semantic Web and of the Asian Conference on the Semantic Web*.

Batista, David S., Rui Silva, Bruno Martins, e Mário J. Silva. 2013. A minwise hashing method for addressing relationship extraction from text. Em *Proceedings of the International Conference on Web Information Systems Engineering*.

Bick, Eckhard. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press Aarhus.

Blessing, Andre e Hinrich Schütze. 2010. Fine-grained geographical relation extraction from wikipedia. Em *Proceedings of the International Conference on Language Resources and Evaluation*.

Branco, António e João Ricardo Silva. 2006. A suite of shallow processing tools for portuguese: Lx-suite. Em *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: Posters and Demonstrations*.

Brin, Sergey. 1999. Extracting patterns and relations from the world wide web. Em *Proceedings of the International Workshop on The World Wide Web and Databases*.

Broder, Andrei. 1997. On the resemblance and containment of documents. Em *Proceedings of the Conference on Compression and Complexity of Sequences*.

Broder, Andrei, Moses Charikar, Alan M. Frieze, e Michael Mitzenmacher. 2000. Min-wise independent permutations. *Journal of Computer and System Sciences*, 60(3).

Bruckschen, Mírian, José Guilherme Camargo de Souza, Renata Vieira, e Sandro Rigo, 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, capítulo Sistema SeRELeP para o reconhecimento de relações entre entidades mencionadas. Linguateca.

Bunescu, Razvan e Raymond Mooney. 2005a. Subsequence kernels for relation extraction. Em *Proceedings of the Annual Conference on Neural Information Processing Systems*.

Bunescu, Razvan C. e Raymond J. Mooney. 2005b. A shortest path dependency kernel for relation extraction. Em *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*.

Cardoso, Nuno. 2008. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. Em *Actas do Encontro do Segundo HAREM*.

Chaves, Marcírio, 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, capítulo Geo-Ontologias para Reconhecimento de Relações Entre Locais: a participação do SEI-Geo no Segundo HAREM. Linguateca.

Chum, Ondrej, James Philbin, e Andrew Zisserman. 2008. Near duplicate image detection: min-hash and TF-IDF weighting. Em *Proceedings of the British Machine Vision Conference*.

Culotta, Aron, Andrew McCallum, e Jonathan Betz. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. Em *Proceedings of*

*the Conference of the North American Chapter of the Association of Computational Linguistics.*

Culotta, Aron e Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Etzioni, Oren, Michele Banko, Stephen Soderland, e Daniel S. Weld. 2008. Open information extraction from the web. *Communication of the ACM*, 51(12):68–74.

Fader, Anthony, Stephen Soderland, e Oren Etzioni. 2011. Identifying relations for open information extraction. Em *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Freitas, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho, e Cristina Mota, 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, capítulo Relaçoes semânticas do ReRelEM: além das entidades no Segundo HAREM. Linguateca.

Gamallo, Pablo, Marcos Garcia, e Santiago Fernández-Lanza. 2012. Dependency-based open information extraction. Em *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP.*

García, Marcos e Pablo Gamallo. 2011. Evaluating various linguistic features on semantic relation extraction. Em *Proocedings of the Conference on Recent Advances in Natural Language Processing.*

GuoDong, Zhou, Su Jian, Zhang Jie, e Zhang Min. 2005. Exploring various knowledge in relation extraction. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Hachey, Ben, Claire Grover, e Richard Tobin. 2012. Datasets for generic relation extraction. *Natural Language Engineering*, 18(1).

Hendrickx, Iris, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó. Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, e Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. Em *Proceedings of the International Workshop on Semantic Evaluation.*

Hoffart, Johannes, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, e Gerhard Weikum. 2012. Kore: keyphrase overlap relatedness for entity disambiguation. Em *Proceedings of the International Conference on Information and Knowledge Management.*

Hoffmann, Raphael, Congle Zhang, e Daniel S Weld. 2010. Learning 5000 relational extractors. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Kambhatla, Nanda. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Posters and Demonstrations.*

Kim, S., J. Yoon, J. Yang, e S. Park. 2010. Walk-weighted subsequence kernels for protein-protein interaction extraction. *BMC Bioinformatics*, 11(107).

Krause, Sebastian, Hong Li, Hans Uszkoreit, e Feiyu Xu. 2012. Large-scale learning of relation-extraction rules with distant supervision from the web. Em *Proceedings of the International Conference on The Semantic Web.*

Li, Ping e Christian König. 2010. b-Bit minwise hashing. Em *Proceedings of the International Conference on World Wide Web.*

Mintz, Mike, Steven Bills, Rion Snow, e Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics and of the International Conference on Natural Language Processing of the Asian Federation of Natural Language Processing.*

Mota, Cristina e Diana Santos. 2008. *Desafios na avaliação onjunta do reconhecimento de entidades mencionadas: O Segundo HAREM.* Linguateca.

Nguyen, Truc-Vien T., Alessandro Moschitti, e Giuseppe Riccardi. 2009. Convolution kernels on constituent, dependency and sequential structures for relation extraction. Em *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Oliveira, Hugo Gonçalo, Hernani Costa, e Paulo Gomes. 2010. Extracção de conhecimento léxico-semântico a partir de resumos da wikipédia. *Actas do II Simpósio de Informática.*

Pantel, Patrick e Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Rajaraman, Anand e Jeffrey Ullman, 2011. *Mining of Massive Datasets*, capítulo 3. Finding Similar Items. Cambridge University Press.

Riedel, Sebastian, Limin Yao, Benjamin M. Marlin, e Andrew McCallum. 2013. Relation extraction with matrix factorization and universal schemas. Em *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics*.

Riedel, Sebastian, Limin Yao, e Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. Em *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*.

Soderland, Stephen e Bhushan Mandhani. 2007. Moving from textual relations to ontologized relations. Em *Proceedings of the AAAI Spring Symposium on Machine Reading*.

Teixeira, Carlos, Arlei Silva, e Wagner Jr. 2012. Min-hash fingerprints for graph kernels: A trade-off among accuracy, efficiency, and compression. *Journal of Information and Data Management*, 3(3).

Wu, Fei e Daniel S Weld. 2010. Open information extraction using wikipedia. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Zelenko, Dmitry, Chinatsu Aone, e Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research*.

Zhao, Shubin e Ralph Grishman. 2005. Extracting relations with integrated information using kernel methods. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Zhou, Guodong e Min Zhang. 2007. Extracting relation information from text documents by exploring various types of knowledge. *Information Processing and Management*, 43(4).

# Clasificación automática del registro lingüístico en textos del español: un análisis contrastivo

**Automatic categorization of Spanish texts into linguistic registers: a contrastive analysis**

John A. Roberto
Universidad de Barcelona
roberto.john@ub.edu

Maria Salamó
Universidad de Barcelona
maria.salamo@ub.edu

M. Antònia Martí
Universidad de Barcelona
amarti@ub.edu

**Resumen**

Las aplicaciones colaborativas como los Sistemas de Recomendación se pueden beneficiar de la clasificación de textos en registros lingüísticos. En primer lugar, el registro lingüístico proporciona información sobre el perfil de los usuarios y sobre el contexto de la recomendación. En segundo lugar, considerar las características de cada tipo de texto puede ayudar a mejorar los métodos actuales de procesamiento de lenguaje natural. En este trabajo contrastamos dos enfoques, uno morfosintáctico y el otro léxico, para categorizar textos por registro en español. Para su evaluación aplicamos 38 algoritmos de aprendizaje automático con los que obtuvimos niveles de precisión superiores al 89 %.

**Palabras clave**

Procesamiento del lenguaje natural, aprendizaje automático, registro lingüístico.

**Abstract**

Collaborative software such as Recommender Systems can benefit from the automatic classification of texts into linguistic registers. First, the linguistic register provides information about the users' profiles and the context of the recommendation. Second, considering the characteristics of each type of text can help to improve existing natural language processing methods. In this paper we contrast two approaches to register categorization for Spanish. The first approach is focused on morphosintactic patterns and the second one on lexical patterns. For the experimental evaluation we tested 38 machine learning algorithms with a precision higher than 89 %.

**Keywords**

Natural language processing, machine learning, linguistic register.

## 1 Introducción

El uso de aplicaciones colaborativas permite a los usuarios crear un tipo de contenidos que por su marcado carácter subjetivo y libre es difícil de tratar computacionalmente. Por esa razón, para que estas aplicaciones funcionen de manera adecuada, deberían garantizar que el tipo de texto que están tratando cumple ciertas condiciones básicas. Por ejemplo, los Sistemas de Recomendación que usan texto libre como mecanismo de reatroalimentación deben garantizar que están procesando opiniones de usuarios reales (*reviews*) y no resúmenes, comentarios o anécdotas sobre determinados productos.

Un sistema recomendador deberá, por tanto, ser capaz de identificar que un texto como el del Ejemplo 1 (tomado de una conocida web de cine) es por su lenguaje, estructura y estilo el comentario de un experto o el resumen del film y no un *review*—entendido este último como un tipo de texto subjetivo que describe la experiencia, el conocimiento y la opinión de un usuario con respecto a un producto (Ricci y Wietsma, 2006).

> Ejemplo **1** *Algo ha pasado, el mundo se va al garete. Una pareja forzosa (José Coronado y Quim Gutiérrez) deberá formar equipo para encontrar a sus seres queridos. Tras las buenas sensaciones dejadas con la aventura americana de "Infectados" (2009), los hermanos Álex y David Pastor regresan a España con "Los últimos días", estupendo drama de ciencia-ficción apocalíptica que les confirma como dos talentos a no dejar escapar. Ahora depende del espectador, claro, y de cómo responda esta crepuscular odisea en nuestro circuito de salas. Sea como fuere, estamos ante uno de los mejores títulos de género nacional de los últimos años. Ni más ni menos.*

Una forma de "validar" los *reviews* es atendiendo a su registro lingüístico. Junto con el género, la temática o el estilo, el registro es un indicador del tipo de texto. En este artículo contrastamos dos aproximaciones para la detección automática del registro lingüístico en español: una aproximación basada en patrones morfosintácticos y otra basada en patrones léxicos. Ambas

aproximaciones tienen en común la simplicidad y reducido coste computacional, dos características indispensables en el diseño de sistemas colaborativos.

En el apartado 2 revisamos las principales aproximaciones empleadas para la clasificación de textos por género o registro. En el apartado 3 describimos los corpus AnCora-ES y Hopinion usados en ésta investigación. En el apartado 4 presentamos las características usadas en cada una de las aproximaciones (morfosintáctica y léxica) y, en el apartado 5, analizamos su rendimiento. Finalmente, en la sección 7 ofrecemos las conclusiones.

## 2 Estado del arte

Los trabajos relacionados con la clasificación automática de textos por registro se pueden agrupar considerando tres factores: la unidad de análisis que hacen servir, la aproximación empleada para la clasificación de los textos y el tipo de rasgos que emplean.

En primer lugar, si consideramos la unidad de análisis tenemos trabajos que se centran en la palabra (Brooke, Wang, y Hirst, 2010), la oración (Lahiri, Mitra, y Lu, 2011) o el documento (Sheikha y Inkpen, 2010). Así, (Brooke, Wang, y Hirst, 2010) exploran diferentes métodos para determinar el nivel de formalidad de ítems léxicos. Su objetivo es clasificar palabras que comparten significado pero no registro, por ejemplo "acquire" *adquirir* (formal) y "snag" *agarrar* (informal). Para conseguir su objetivo, los autores emplean la longitud de las palabras mediante *FS score* (Simple Formality Measure), LSA (Latent Semantic Analysis) y un método híbrido que combina los dos anteriores. Por su parte, (Lahiri, Mitra, y Lu, 2011) analizan el grado de formalidad a nivel de la oración mediante el cálculo del F-score (Formality Score) y evaluando el grado de acuerdo entre anotadores a partir de los coeficientes Kappa y Jaccard. Por último, los trabajos que vienen a continuación tratan el problema del registro a nivel de documento.

En segundo lugar, las aproximaciones empleadas para la clasificación de textos por registro o género se basan en el análisis de patrones lingüísticos y en el uso de técnicas de aprendizaje automático. La propuesta más representativa basada en el análisis de patrones lingüísticos es la metodología del análisis multidimensional (MDA) de Biber (1988). En su trabajo, Biber aplica 67 rasgos lingüísticos para identificar 21 géneros del inglés hablado y escrito. Biber determina los rasgos que concurren en un mismo género mediante técnicas estadísticas de análisis de factores. En la misma línea, (Tribble, 1999) propone un método menos complejo que el MDA consistente en caracterizar los textos a partir de la detección de palabras clave. Tribble extrae listas de palabras de forma automática y las compara con un corpus de referencia para obtener las palabras más relevantes dentro de un género o registro. El método de Tribble fue contrastado con el de Biber por (Xiao y McEnery, 2005) obteniendo resultados similares.

Más recientes pero también más escasas, son las aproximaciones basadas en técnicas de aprendizaje automático (especialmente para el español). En cuanto al aprendizaje supervisado, (Sharoff, Zhili, y Katja, 2010) emplean el algoritmo SVM (*Support Vector Machines*), un modelo basado en trigramas de etiquetas POS y las anotaciones del corpus Brown (Francis y Kucera, 1979) para la detección del género en textos de la Web. Según los autores, el uso de rasgos léxicos es más efectivo para detectar el género de un documento que la información basada en Part-Of-Speech. En (Gries, Newman, y Shaoul, 2009) se presenta un algoritmo de clústering aglomerativo jerárquico basado en n-gramas para detectar registros en textos provenientes de dos corpus diferentes: el BNC-Baby (*British National Corpus Baby*) y el ICE-GB (*British Component of the International Corpus of English*).

En tercer lugar, como comentamos, las investigaciones en clasificación automática de textos por registro se pueden agrupar considerando el tipo de rasgos que hacen servir. Son características las bolsas de palabras (zu Eissen y Stein, 2004), n-gramas de caracteres (Mason, Shepherd, y Duffy, 2009; Kanaris y Stamatatos, 2007) y Part-Of-Speech (Santini, 2007), así como el uso de etiquetas HTML para analizar el metacontenido de las páginas (Boese y Howe, 2005). Más reciente es el uso de rasgos léxico-gramaticales (Sheikha y Inkpen, 2010) como pueden ser la frecuencia de interjecciones, palabras formales e informales, uso de la forma pasiva, etc. En estas aproximaciones se suele evaluar el rendimiento y la utilidad de tales rasgos mediante la aplicación de técnicas de aprendizaje automático como las descritas.

En español tenemos el trabajo de (Mosquera y Moreda, 2011). Los autores de esa investigación identifican grados de informalidad en textos de la Web 2.0 usando un algoritmo de "hard-clustering" (K-Means) y un conjunto de 19 características léxico-gramaticales. Las conclusiones a las que llegan son positivas aunque señalan la necesidad de añadir nuevas características.

La principal novedad de nuestra propuesta en relación con las anteriores investigaciones consiste en hacer un análisis exhaustivo, con 38 algoritmos de aprendizaje automático y diversas técnicas de selección de atributos, de dos aproximaciones diferentes para la clasificación de textos según su registro. A diferencia de Mosquera, nosotros usamos aprendizaje supervisado para categorizar los textos en dos grandes grupos (formal e informal) y esto lo hacemos distinguiendo entre características morfosintácticas y léxicas. En cuanto a estas últimas, otro aporte a destacar es la incorporación de diferentes métricas de la riqueza léxica para la detección del registro lingüístico.

## 3 Datos

En este artículo se ha utilizado un subconjunto de 3270 textos tomados de los corpus AnCora-ES y Hopinion. Usamos estos dos corpus ya que representan registros opuestos del español actual:

- AnCora-ES es un corpus del español formal constituido principalmente por artículos periodísticos. AnCora-ES está anotado con diferentes tipos de información lingüística, por ejemplo, *Part of Speech*, estructura argumental, papeles temáticos, correferencia, entre otros.

- Hopinion es un corpus del español coloquial constituido por opiniones de hoteles descargadas de la web de TripAdvisor. Hopinion está anotado con información morfosintáctica que ha sido revisada manualmente por un grupo de lingüistas.

El Cuadro 1 describe las características principales del subconjunto de datos utilizado. Por simplicidad, en adelante nos referiremos a ambos subconjuntos como corpus AnCora y corpus Hopinion.

| Característica | Hopinion | AnCora-Es |
|---|---|---|
| Número de textos | 1635 | 1635 |
| Total palabras | 206.812 | 443.380 |
| Promedio de palabras por texto | 126.49 | 271.18 |
| Fuente de datos | TripAdvisor | El Periódico Agencia EFE |
| Registro asociado | Colloquial | Formal |

Cuadro 1: Características de los corpus.

## 4 Aproximaciones

En esta sección se describen las dos aproximaciones contrastadas en nuestro estudio.

### 4.1 Aproximación basada en patrones morfosintácticos

Esta aproximación utiliza una serie de características morfosintácticas para la clasificación automática de los textos según su registro lingüístico. Las características que hemos seleccionado, once en total (ver Cuadro 2), representan cinco de las principales manifestaciones lingüísticas descritas en los estudios sobre el español coloquial:

- Sintaxis concatenada (SXC): consiste en la acumulación de enunciados producto de la ausencia de planificación en la producción del mensaje (Narbona, 1989).

- Elipsis (ELP): es la omisión de elementos lingüísticos que se presuponen a partir de entidades que se hallan presentes en el contexto discursivo. Una forma básica pero muy frecuente de representar tales omisiones en el lenguaje coloquial, es mediante el uso de los puntos suspensivos.

- Redundancia (RED): consiste en duplicar, de manera exacta o aproximada, algunas partes del discurso (Tannen, 1989).

- Deixis (DXS): es un recurso para la cohesión textual que el hablante utiliza para introducir las entidades o referentes del contexto situacional en el discurso.

- Riqueza léxica (RL): son diferentes métricas que se utilizan para conocer la competencia léxica de un hablante (Read, 2005).

| N° | Característica | Manifest. |
|---|---|---|
| 1 | Densidad léxica | RL |
| 2 | Signos de puntuación | SXC |
| 3 | Co-occurencia de palabras | RED |
| 4 | Conjunciones coordinantes | SXC |
| 5 | Conjunciones subordinantes | SXC |
| 6 | Pronombres personales y demostrativos | DXS |
| 7 | Puntos suspensivos | ELP |
| 8 | Interjecciones | PAR |
| 9 | Repetición de vocales y consonantes | PAR |
| 10 | Oraciones consecutivas[1] | INT |
| 11 | Variación léxica (TTR) | RL |

Cuadro 2: Características morfosintácticas evaluadas en los textos.

Para calcular las frecuencias de estas once características, los textos se han etiquetado con Part-Of-Speech y lema. La riqueza léxica se ha obtenido mediante el cálculo de la densidad (ver Ecuación 1) y la variación[2] (ver Ecuación 2) léxicas.

$$DL = \frac{palabras\ léxicas}{total\ palabras} \times 100 \quad (1)$$

$$VL = \frac{palabras\ diferentes}{total\ palabras} \times 100 \quad (2)$$

### 4.2 Aproximación basada en patrones léxicos

Para obtener el registro lingüístico de los textos, esta aproximación se basa en la detección de términos informales y de emoción conjuntamente con el cálculo de la riqueza léxica.

Por una parte, los términos con un uso informal se obtuvieron consultando las versiones en línea de Wikcionario[3] y TheFreeDictionary[4]. Las marcas de uso que se hicieron servir para reconocer dichos términos fueron: "Coloquial", "Despectivo", "Malsonante", "Familiar", "Informal", "Peyorativo" y "Vulgar". De otro lado, la detección de los términos de emoción se efectuó mediante el *Spanish Emotion Lexicon* (Sidorov et al., 2012), un recurso léxico creado de forma totalmente manual por investigadores del Instituto Politécnico Nacional de México.

En la Figura 1 tenemos dos ejemplos del formato (XML) y el tipo de información con que se ha anotado cada una de las palabras en los corpus Ancora y Hopinion a partir de los recursos señalados.

Conjuntamente con los atributos "registro" (*register*) y "emoción" (*emotion*) hemos usado nueve métricas de la riqueza léxica[5] (Roberto, Martí, y Salamó, 2012): densidad léxica (ver Ecuación 1), type token ratio (ver Ecuación 2), sofisticación léxica (ver Ecuación 3), perfíl de frecuencia léxica (ver Ecuación 4), $a^2$ (ver Ecuación 5), índice de Uber (ver Ecuación 6), Z de Zipf (ver Ecuación 7), variación de palabras léxicas (ver Ecuación 8) y variación modal (ver Ecuación 9).

[1]El patrón que usamos para identificar los esquemas oracionales consecutivos es: *[intensificador: tanto, tan, tal, etc.] + [nombre OR adjetivo OR adverbio] + [que]*

[2]Concretamente *type-token ratio*.

[3]http://es.wiktionary.org/

[4]http://es.thefreedictionary.com/

[5]Incluidas las dos de la aproximación morfosintáctica.

```
<wd lemma="cabrear"
    register="coloquial"
    emotion="enojo"
    source="ancora" />

<wd lemma="acojonante"
    register="coloquial"
    emotion="sorpresa"
    source="hopinion" />
```

Figura 1: Ejemplos de palabras anotadas con registro y emoción.

$$SL = \frac{N_{slex}}{N_{lex}} \quad (3)$$

$$PFL = \frac{T_s}{T} \quad (4)$$

$$a^2 = \frac{\log N - \log T}{\log^2 N} \quad (5)$$

$$IU = \frac{(\log N)^2}{logN - logT} \quad (6)$$

$$ZIPF = \frac{Z \times N \times \log(N/Z)}{(N - Z)\log(p \times Z)} \quad (7)$$

$$VPL = \frac{T_{lex}}{N_{lex}} \quad (8)$$

$$VM = \frac{T_a + T_r}{N_{lex}} \quad (9)$$

**Note:** $N$ (tokens), $T$ (types), $lex$ (unidades léxicas), $s$ (unidades sofisticadas), $p$ (token más frecuente dividido por la longitud del texto) y $Z$ (una medida de la riqueza léxica, en este caso $Z = TTR$).

Para obtener estos valores hemos usado la herramienta para el Análisis de Textos de Opinión en lenguaje natural (ATOp) (Queral, 2013). ATOp es una plataforma en Java que estamos desarrollando como parte de nuestro trabajo en minería de opiniones (ver Figura 2).

## 5 Evaluación y resultados

En esta sección presentamos los resultados obtenidos al aplicar las dos aproximaciones para la detección automática del registro lingüístico descritas en la sección anterior.

En los experimentos empleamos técnicas de aprendizaje supervisado el cual nos permite predecir la clase a la que pertenece un determinado objeto a partir de una serie de ejemplos de entrenamiento. En nuestro caso, el objetivo es predecir si un texto $x$ pertenece a la clase formal (AnCora) o coloquial (Hopinion) basándonos en

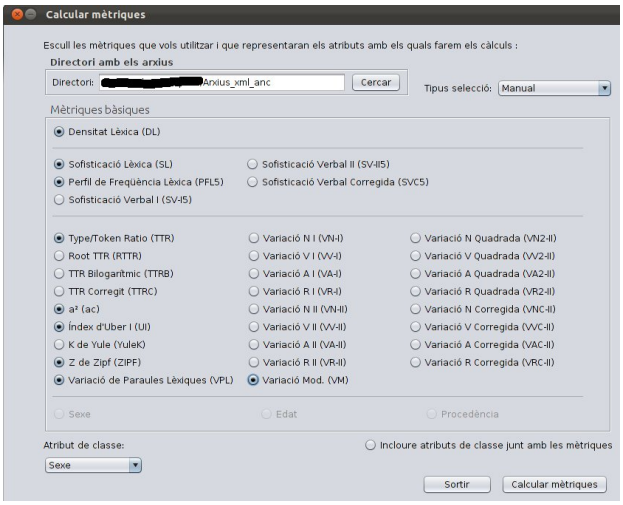


Figura 2: Selección de las métricas de riqueza léxica en ATOp.

las carecterísticas o atributos enumerados en la sección 4.

Como herramienta de análisis se ha usado Weka (Witten y Frank, 2000). De esta herramienta se han seleccionado 38 conocidos algoritmos de aprendizaje automático supervisado, los cuales se pueden dividir en 5 grandes categorías, ver Cuadro 3. En dicho cuadro, la primera columna describe la categoría y la segunda los algoritmos usados para la evaluación de esa categoría.

Todos los experimentos tienen la misma configuración. Se ha recopilado un fichero de datos donde se describen todos los textos en función de un conjunto de atributos, éstos se corresponden con las características descritas anteriormente (en la Sección 4) y como atributo de clase se ha definido el registro "coloquial" o "formal". La experimentación se ha realizado con validación cruzada[6] (*ten-fold cross-validation*). El resultado de los clasificadores se da en términos de su precisión (*Prediction Accuracy*), es decir, el porcentaje de instancias que fueron correctamente clasificadas.

Con el fin de determinar los atributos que tienen más peso, hemos aplicado varios métodos de selección de atributos. Los métodos de evaluación y de búsqueda usados en la selección supervisada se enumeran en el Cuadro 4. Los métodos de selección de atributos reducen el número de variables, seleccionando el mejor subconjunto de características del conjunto de inicial.

[6]La validación cruzada es una técnica para la evaluación experimental en la que los datos disponibles se dividen aleatoriamente en un conjunto de entrenamiento y un conjunto de test.

| Categoría | Algoritmo de Aprendizaje |
|---|---|
| Bayes | BayesNet, BayesianLogisticRegression, ComplementNaiveBayes, DMNBtext, NaiveBayes, NaiveBayesMultinomial, NaiveBayesMultinomialUpdateable, NaiveBayesSimple, NaiveBayesUpdateable |
| Lazy | IB1, IBk, KStar, LWL |
| Misc | HyperPipes, VFI |
| Rules | ConjunctiveRule, DTNB, DecisionTable, JRip, NNge, OneR, PART, Ridor, ZeroR |
| Trees | ADTree, BFTree, DecisionStump, FT, J48, J48graft, LADTree, LMT, NBTree, REPTree, RandomForest, RandomTree, SimpleCart, lmt.LogisticBase |

Cuadro 3: Listado de los algoritmos utilizados. Como se puede observar Weka permite elegir entre múltiples algoritmos de clasificación, distribuidos en cinco categorías.

| **Métodos de evaluación** |
|---|
| CfsSubsetEval, ChiSquaredAttributeEval, ConsistencySubsetEval, FilteredAttributeEval, FilteredSubsetEval, GainRatioAttributeEval, InfoGainAttributeEval, LatentSemanticAnalysis, OneRAttributeEval, PrincipalComponents, ReliefFAttributeEval, SVMAttributeEval, WrapperSubsetEval |
| **Métodos de búsqueda** |
| BestFirst, ExhaustiveSearch, GeneticSearch, GreedyStepwise, LinearForwardSelection, RandomSearch, RankSearch, Ranker, ScatterSearchV1, SubsetSizeForwardSelection |

Cuadro 4: Métodos de selección de atributos usados por los clasificadores en la experimentación y divididos en dos características: el método de evaluación utilizado y el método de búsqueda.

### 5.1 Clasificación mediante patrones morfosintácticos

En este apartado describimos los niveles de precisión alcanzados al predecir el registro lingüístico de los textos en los corpus Hopinion y Ancora, usando las once características morfosintácticas enumeradas en la sección 4.1.

Aplicando la configuración detallada al principio de esta sección entrenamos un total de 14.400 clasificadores[7]. En promedio, los clasificadores que usan algún método de selección de atributos funcionan mejor que los que no los usan (ver Figura 3).

[7]La combinación de algunos métodos de evaluación y de búsqueda no son posibles en Weka.

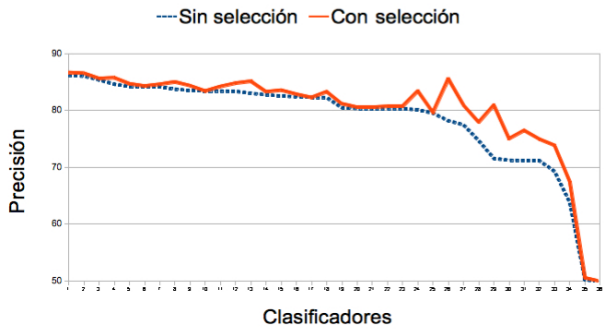


Figura 3: Precisión con y sin selección de atributos. Usando la selección de atributos se consigue reducir el número de características, mientras que la precisión mejora o se mantiene.

El clasificador que obtuvo un nivel de precisión más elevado (89 %) fue *trees.RandomForest* con *CfsSubsetEval* como método de evaluación y de búsqueda *Ranker*. Random Forests es una técnica de agregación que incorpora aleatoriedad en la construcción de cada clasificador para mejorar la precisión. El método *CfsSubsetEval* considera el valor predictivo individual de cada atributo seleccionando aquellos que estén altamente correlacionados con la clase y tengan entre ellos baja intercorrelación. *Ranker*, por su parte, devuelve una lista ordenada de los atributos según su calidad.

Para identificar los atributos con mayor valor predictivo, en el Cuadro 5 hemos contrastado el rendimiento promedio de los clasificadores cuando usan un determinado atributo ($A_x$) y cuando dejan de usarlo ($\neg A_x$). De esta manera, el valor *Diferencia* determinará el impacto que tiene la omisión del atributo $x$ a nivel de la precisión.

| | *Promedios* | | | |
|---|---|---|---|---|
| $A$ | $A_x$ | $\neg A_x$ | *Diferencia* | |
| **1** | 78.3 % | 65.5 % | 12.8 | |
| **2** | 78.5 % | 76.1 % | 2.4 | |
| **3** | 78.1 % | 72.7 % | 5.4 | |
| **4** | 78.6 % | 62.2 % | 16.4 | • |
| **5** | 78.6 % | 76.3 % | 2.3 | |
| **6** | 78.6 % | 62.2 % | 16.4 | • |
| **7** | 78.6 % | 65.7 % | 12.9 | |
| **8** | 78.5 % | 71.3 % | 7.2 | |
| **9** | 78.6 % | 76.2 % | 2.4 | |
| **10** | 78.6 % | 76.1 % | 2.5 | |
| **11** | 78.6 % | 65.9 % | 12.7 | |

Cuadro 5: Atributos morfosintácticos con mayor valor predictivo.

Los atributos más informativos son el 4 y el 6, es decir, conjunciones coordinantes y pronombres personales y demostrativos (DXS). Los menos informativos son el 2 y el 9 (signos de puntuación y repetición de vocales y consonantes).

Finalmente, para conocer el rendimiento de los clasificadores, en la Figura 4 relacionamos la precisión de cada algoritmo de clasificación con el número de atributos seleccionados. Como se observa, es posible obtener niveles de precisión adecuados con solo 6 atributos. De manera específica, la precisión media de los clasificadores que usan seis atributos es de 85.6 %, siendo *trees.LMT* (*Logistic Model Tree*) el clasificador que obtuvo el valor más alto: 85.9 % (también con seis atributos).

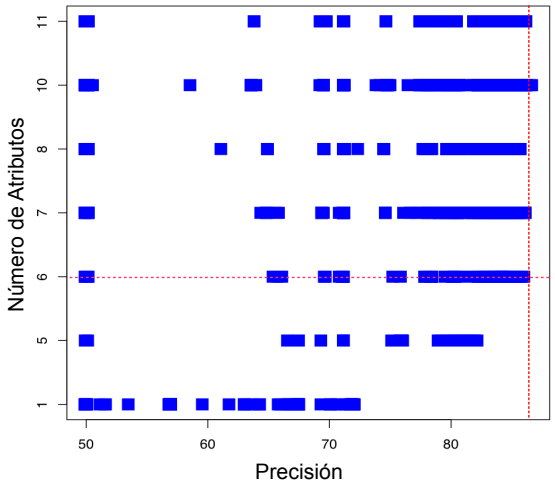


Figura 4: Rendimiento: número de atributos morfosintácticos versus precisión.

### 5.2 Clasificación mediante patrones léxicos

En este apartado describimos los niveles de precisión alcanzados al predecir el registro lingüístico de los textos en los corpus Hopinion y Ancora, usando las once características léxicas enumeradas en la sección 4.2.

En total entrenamos 13.768 clasificadores. A diferencia de lo que sucede con la clasificación mediante patrones morfositácticos, los clasificadores que usan algún método de selección de atributos y los que no los usan se comportan de manera similar. En ambos casos obtuvimos una precisión promedio del 84 %.

El clasificador que obtuvo un nivel de precisión más elevado (93.8 %) usa selección de atributos: *trees.RandomForest* con *ConsistencySubsetEval* como método de evaluación y de búsqueda *RankSearch*. El método *ConsistencySubsetEval* mide la consistencia de un subconjunto de atributos en términos de las clases. *RankSearch* ordena los atributos utilizando un evaluador individual o de conjuntos y crea un ranking de subconjuntos prometedores.

Debido a que nueve de los once atributos tiene que ver con la riqueza léxica y dos con el uso de términos coloquiales y de emoción, en la Figura 5 (ver Apéndice) presentamos la relación que hay entre estos tres tipos de atributos y la precisión. Como se puede observar, la mediana más alta la tienen los clasificadores que usan los tres tipos de atributos (RIQ+USO+EMO), luego están los que utilizan "riqueza" conjuntamente con "uso" (RIQ+USO), seguidos por los clasificadores que utilizan solo el atributo "emoción" (EMO) o "riqueza" (RIQ). Ningún clasificador emplea el atributo USO de forma aislada.

Adicionalmente, en la Figura 6 (ver Apéndice) presentamos los atributos que han seleccionado los 13.768 clasificadores como los más relevantes. En total tenemos seis atributos, el de emoción, el del uso coloquial y cuatro de riqueza léxica. En este último caso la densidad léxica (DL) ha dado mejores resultados.

Por último, en la Figura 7 (ver Apéndice) tenemos el número de atributos usados por los clasificadores para obtener los distintos niveles de precisión. Los clasificadores que tienen un mejor rendimiento usan siete atributos. Estos clasificadores presentan precisiones máximas superiores al 90 % y sus medianas están entre el 80 % y el 90 %.

## 6 Discusión

Nuestros experimentos indican que es posible usar el registro lingüístico para clasificar textos del español de manera fiable. El empleo de patrones léxicos ha sido más efectivo que la aproximación morfosintáctica, tanto a nivel de precisión como al número de atributos empleados. A la misma conclusión llegan (Sharoff, Zhili, y Katja, 2010) en su estudio sobre clasificación de textos por género para el inglés. Si bien, tal como se comenta en el mismo estudio, la eficiencia de los patrones léxicos puede estar relacionada con su capacidad para predecir el dominio antes que el género o el registro de los textos, éste no es nuestro caso ya que los rasgos empleados (riqueza léxica, términos coloquiales y de emoción) son independientes del dominio.

## 7 Conclusiones

En este artículo hemos contrastado dos aproximaciones para la clasificación del registro lingüístico en textos del español. La precisión de base obtenida con la aproximación morfosintáctica fue del 89 % (con diez atributos) y con la léxica del 93.8 % (con siete atributos).

Los atributos más informativos son, en la aproximación morfosintáctica, las conjunciones coordinantes, los pronombres personales y los demostrativos. En la aproximación léxica son la densidad léxica, los términos de emoción y los de uso coloquial. Este último siempre aparece correlacionado con otros atributos.

## Agradecimientos

Este trabajo ha sido posible gracias los proyectos DIANA (DIscourse ANAlysis for knowledge understanding, TIN2012-38603) y TIN2009-14404-CO2 del Ministerio de Ciencia e Innovación, así como a la beca FI 2010FI_B 00521 de la Generalitat de Catalunya.

## Bibliografía

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge, UK.

Boese, Elizabeth S. y Adele E. Howe. 2005. Effects of web document evolution on genre classification. En *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, páginas 632–639, New York, NY, USA. ACM.

Brooke, Julian, Tong Wang, y Graeme Hirst. 2010. Automatic acquisition of lexical formality. En *In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING*.

Francis, W. y H. Kucera. 1979. Brown corpus. Text, Department of Linguistics, Brown University.

Gries, Stefan, John Newman, y Cyrus Shaoul. 2009. N-grams and the clustering of genres. *ELR Journal*, 5:1–13.

Kanaris, Ioannis y Efstathios Stamatatos. 2007. Webpage genre identification using variable-length character n-grams. En *In Proc. of the 19th IEEE Int. Conf. on Tools with Artificial Intelligence, v.2*, páginas 3–10.

Lahiri, Shibamouli, Prasenjit Mitra, y Xiaofei Lu. 2011. Informality judgment at sentence level and experiments with formality score. En *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part II*, CICLing'11, páginas 446–457, Berlin, Heidelberg. Springer-Verlag.

Mason, J., M. Shepherd, y J. Duffy. 2009. An n-gram based approach to automatically identifying web page genre. En *Proceedings of the 42nd Hawaii International Conference on System Sciences*, HICSS '09, páginas 1–10, Washington, DC, USA. IEEE Computer Society.

Mosquera, A. y P. Moreda. 2011. Caracterización de niveles de informalidad en textos de la web 2.0. *Procesamiento del Lenguaje Natural*, 47:171–177.

Narbona, Antonio. 1989. *Sintaxis española: nuevos y viejos enfoques*. Ariel.

zu Eissen, Sven Meyer y Benno Stein. 2004. Genre classification of web pages: User study and feasibility analysis. En *IN: BIUNDO S., FRUHWIRTH T., PALM G. (EDS.): ADVANCES IN ARTIFICIAL INTELLIGENCE*, páginas 256–269. Springer.

Queral, Bakary Singateh. 2013. Plataforma en java per l'anàlisi de textos d'opinió en llenguatge natural (atop). Master's thesis, Universitat de Barcelona. Facultat de Matemàtiques.

Read, J. 2005. *Assessing vocabulary*. Cambridge University Press, 5 edición.

Ricci y Wietsma. 2006. Product reviews in travel decision making. *Proceeding of Information and Communication Technologies in Tourism (ENTER)*, páginas 296–307.

Roberto, J., M. Martí, y M. Salamó. 2012. Análisis de la riqueza léxica en el contexto de la clasificación de atributos demográficos latentes. *Procesamiento de Lenguaje Natural*, 1(48):97–104.

Santini, Marina. 2007. *Automatic Identification of Genre in Web Pages*. Ph.D. tesis, University of Brighton.

Sharoff, Serge, Wu Zhili, y Markert Katja. 2010. The web library of babel: evaluating genre collections. *Proceedings of Seventh International Conference on Language Resources and Evaluation (LREC10)*, páginas 3063–3070.

Sheikha, Abu y Diana Inkpen. 2010. Automatic classification of documents by formality. En *International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, páginas 1–5.

Sidorov, G., S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. Suárez-Guerra, A. Treviño, y J. Gordon. 2012. Empirical study of opinion mining in spanish tweets. *LNAI*, páginas 7629–7630.

Tannen, D. 1989. Repetition in conversation: Towards a poetics of talk. En D. Tannen, editor, *Talking Voices. Repetition, dialogue and imagery in conversational discourse*. Cambridge, CUP.

Tribble, Christopher. 1999. *Writing Difficult Texts*. Ph.D. tesis, Lancaster University.

Witten, I. y E. Frank. 2000. *DataMining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann Publishers.

Xiao, Zhonghua y Anthony McEnery. 2005. Two approaches to genre analysis. three genres in modern american english. *Journal of English Linguistics*, 33(3):62–82.

## Apéndice

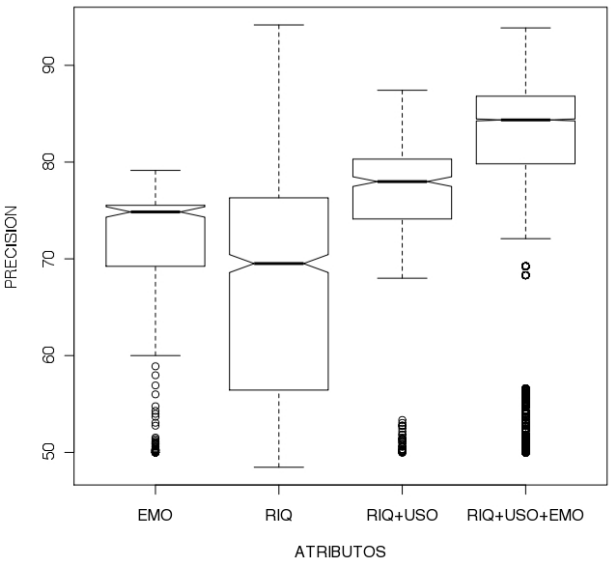


Figura 5: Precisión en relación con los tipos de atributos léxicos "emoción", "riqueza" y "uso".

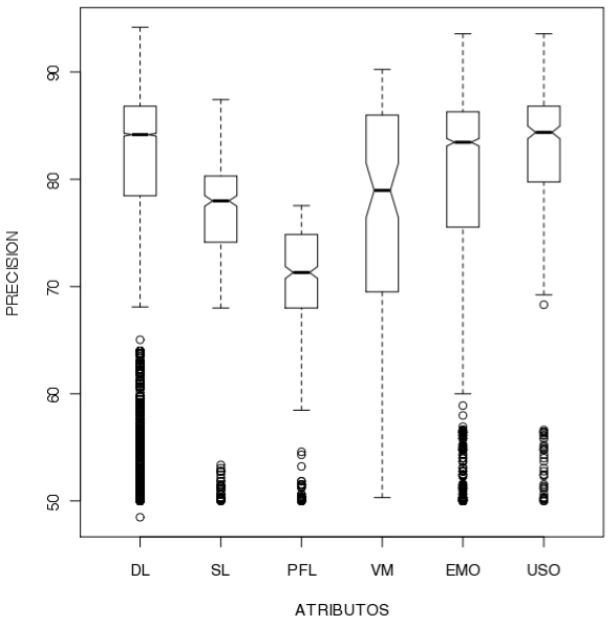


Figura 6: Mejores atributos léxicos seleccionados por los clasificadores.

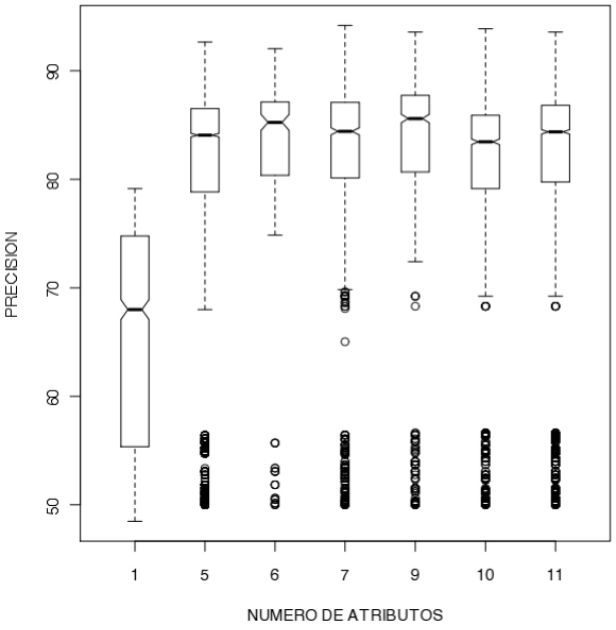


Figura 7: Precisión en relación con el número de atributos léxicos utilizados por los clasificadores.

Volume 5, Número 1- Julho 2013

linguaMATICA

ISSN: 1647-0818

## Dossier

Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual Central Repository
*Aitor Gonzalez-Agirre e German Rigau*

## Artigos de Investigação

Un método de análisis de lenguaje tipo SMS para el castellano
*Andrés Alfonso Caurcel Díaz, José María Gómez Hidalgo e Yovan Iñiguez del Rio*

Extracção de relações semânticas de textos em português explorando a DBpédia e a Wikipédia
*David S. Batista, David Forte, Rui SIlva, Bruno Martins e Mário J. Silva*

Clasificación automática del registro lingüístico en textos del español: un análisis contrastivo
*John A. Roberto, Maria Salamó e M. Antònia Martí*

http://linguamatica.com