Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 69-76 recibido 15-04-14 revisado 01-07-14 aceptado 02-07-14

Descripción y Evaluación de un Sistema de Extracción de Definiciones para el Catalán

Description and Evaluation of a Definition Extraction System for Catalan

Luis Espinosa-Anke, Horacio Saggion
Universitat Pompeu Fabra
C/ Tànger, 122-134, 4a Planta
{luis.espinosa, horacio.saggion}@upf.edu

Resumen: La extracción automática de definiciones (ED) es una tarea que consiste en identificar definiciones en texto. Este artículo presenta un método para la identificación de definiciones para el catalán en el dominio enciclopédico, tomando como corpora para entrenamiento y evaluación una colección de documentos de la Wikipedia en catalán (Viquipèdia). El corpus de evaluación ha sido validado manualmente. El sistema consiste en un algoritmo de clasificación supervisado basado en Conditional Random Fields. Además de los habituales rasgos lingüísticos, se introducen rasgos que explotan la frecuencia de palabras en dominios generales y específicos, en definiciones y oraciones no definitorias, y en posición de definiendum (el término que se define) y de definiens (el clúster de palabras que define el definiendum). Los resultados obtenidos son prometedores, y sugieren que la combinación de rasgos lingüísticos y estadísticos juegan un papel importante en el desarrollo de sistemas ED para lenguas minoritarias.
Palabras clave: Extracción de Definiciones, Extracción de Información, Lexicografía Computacional

Abstract: Automatic Definition Extraction (DE) consists of identifying definitions in naturally-occurring text. This paper presents a method for the identification of definitions in Catalan in the encyclopedic domain. The train and test corpora come from the Catalan Wikipedia (Viquipèdia). The test set has been manually validated. We approach the task as a supervised classification problem, using the Conditional Random Fields algorithm. In addition to the common linguistic features, we introduce features that exploit the frequency of a word in general and specific domains, in definitional and non-definitional sentences, and in definiendum (term to be defined) and definiens (cluster of words that defines the definiendum) position. We obtain promising results that suggest that combining linguistic and statistical features can prove useful for developing DE systems for under-resourced languages.
Keywords: Extracción de Definiciones, Extracción de Información, Conditional Random Fields, Wikipedia

1 Introducción

Las enciclopedias y bases de datos terminológicas son bases de conocimiento de gran importancia para establecer relaciones semánticas entre distintos conceptos. No obstante, el desarrollo manual de estos recursos es habitualmente costoso y lento (Bontas y Mochol, 2005). La Extracción automática de Definiciones (ED), entendida como la tarea de identificar automáticamente definiciones en una producción lingüística natural, puede jugar un papel importante en este contexto. Existe un creciente interés en la ED en los campos del Procesamiento del Lenguaje Natural, la Lingüística Computacional y la Lexicografía Computacional. Además, trabajo previo ha demostrado su potencial para el desarrollo automático de glosarios (Muresan y Klavans, 2002; Park, Byrd y Boguraev, 2002), bases de datos léxicas (Nakamura y Nagao, 1988) herramientas de búsqueda de respuestas (Saggion y Gaizauskas, 2004; Cui, Kan y Chua, 2005), como apoyo para aplicaciones terminológicas (Meyer, 2001; Sierra

ISSN 1135-5948

et al., 2006), o el desarrollo de aplicaciones de aprendizaje en línea (Westerhout y Monachesi, 2007; Espinosa-Anke, 2013).

Este artículo presenta un trabajo orientado a la ED para el catalán basado en aprendizaje automático. En primer lugar, se describe la compilación y creación de un corpus extraído de Viquipèdia[1] (la Wikipedia catalana). Partimos de la idea de que un concepto monosémico con una entrada en Viquipèdia será definido en la primera oración de su artículo. Además, consideramos como no definiciones otras oraciones del mismo artículo en el que el término también aparece, pues si bien aportan información relevante para el término definido, ésta no es factual e independiente del contexto en el que se enuncia. De hecho, en muchos casos son "definiciones falsas sintácticamente plausibles" (Navigli, Velardi y Ruiz-Martínez, 2010), es decir, oraciones que muestran un comportamiento sintáctico muy similar al de una definición.

Nuestro corpus de entrenamiento consiste en un subcorpus de la rama en catalán de Wikicorpus (Reese et al., 2010), mientras que el corpus de test consiste en un subconjunto de una versión en catalán del WCL dataset (Navigli, Velardi y Ruiz-Martínez, 2010). El corpus de entrenamiento consiste en 10375 definiciones y 8010 no definiciones. Por otra parte, el corpus de test incluye 1407 no definiciones y 2796 definiciones.

El proceso de aprendizaje automático se basa en el uso del algoritmo Conditional Random Fields (CRF) (Lafferty, McCallum, y Pereira, 2001), específicamente el *toolkit* CRFsuite (Okazaki, 2007). CRF es un algoritmo de etiquetado secuencial que permite la incorporación de rasgos de observaciones y transiciones no sólo adyacentes, sino también de larga distancia, previos y posteriores a la observación actual. Es apropiado para nuestra tarea ya que consideramos la ED como una tarea de etiquetado secuencial, en la que cada palabra puede estar al principio, dentro o fuera de una definición.

El resto del artículo se estructura de la siguiente manera: La Sección 2 repasa trabajo anterior en el campo de la clasificación de definiciones, además de la tarea específica de ED. A continuación, la Sección 3 describe el proceso de compilación de los corpora utilizados en este estudio, además de los rasgos aplicados en el modelado de datos y su justificación. La Sección 4 muestra los resultados obtenidos con distintas configuraciones. Finalmente, las secciones 5 y 6 ofrecen un a visión general y resumida del trabajo descrito en este artículo, y señalan futuras líneas de investigación en el ámbito de ED para lenguas minoritarias, respectivamente.

2 Estado de la Cuestión

La Extracción automática de Definiciones (ED) es una tarea que consiste en identificar oraciones que incluyen información definitoria (Navigli y Velardi, 2010). El procesado automático de textos con fines terminográficos puede constituir una herramienta que facilite la elaboración de glosarios o diccionarios especializados, bases de datos de conocimiento léxico o bien para la elaboración de ontologías (Alarcón, 2009).

El estudio de la relación entre un concepto unívoco y monosémico y su definición se remonta al modelo aristotélico de definición, el modelo *genus et differentia*, en el que un término es definido mencionando su género próximo además de un conjunto de características particulares. Generalmente, se conoce al término definido como *definiendum*, y al clúster de palabras que le define, *definiens*.

Partiendo de este modelo, el concepto de definición se ha ido elaborando con la inclusión de distintos factores. Por ejemplo, Trimble (1985) propone una clasificación basada en el grado de formalidad y la información que se transmite, Auger y Knecht (1997) se refieren a *enunciados de interés definitorio* como la inclusión de aspectos como el sentido o contexto de uso, entre otros, y su relación con un concepto o idea específicos. Un estudio destacable en esta línea es el de Meyer (2001), que acuña el concepto de *contexto rico en conocimiento* (CRC), definido como "naturally occurring utterances that explicitly describe attributes of domain-specific concepts or semantic relations holding between them at a certain point in time, in a manner that is likely to help the reader of the context understand the concept in question". Destacamos en el ámbito del estudio y extracción automática de CRCs el trabajo de Feliu y Cabré (2002), en el que se propone un sistema basado en reglas para la identificación y clasificación automática de CRCs para el catalán, basado en la combinación de patro-

[1] https://ca.wikipedia.org/

nes léxico-sintácticos y medidas estadísticas para evaluar la prominencia de un candidato a término en un contexto concreto. Por otra parte, en ED existe una creciente tendencia a aplicar algoritmos de aprendizaje automático (Del Gaudio, Batista, y Branco, 2013).

A continuación se describe el método empleado para la construcción del corpus utilizado para entrenar y evaluar nuestro sistema, así como la descripción de los rasgos lingüísticos y estadísticos utilizados para el modelado de datos.

3 Método

En esta sección se describen los datos utilizados para entrenar y evaluar el sistema de ED y los rasgos utilizados en el modelado de los datos.

3.1 Los datasets

Tomamos como base un corpus de definiciones y oraciones no definitorias sobre un determinado término (Navigli, Velardi y Ruiz-Martínez, 2010). A partir de él, se realiza un proceso de mapeo entre aquellos términos que aparecen en el corpus original y su equivalente en Viquipèdia. Se aplican una serie de reglas para evitar ruido y evitar mapeos en blanco dado que existen términos en Wikipedia sin una entrada equivalente en otro idioma, en este caso el catalán.

A continuación se muestran dos ejemplos de oraciones definitorias y no definitorias en catalán, presentes en nuestro corpus, para el término *iot* (yate).

- **Def**: Un iot és una embarcació d'esbarjo o esportiva propulsada a vela o a motor amb coberta i amb cabina per a viure-hi.
- **Def**: *Un yate es una embarcación de recreo o deportiva propulsada a vela o a motor con cubierta y cabina para vivir.*
- **Nodef**: Tot i això la majoria de iots a vela privats solen tenir una eslora de 7 a 14m, ja que el seu cost augmenta ràpidament en proporció a l'eslora.
- **Nodef**: *Sin embargo, la mayoría de yates a vela privados suelen tener una eslora de 7 a 14m, ya que su coste aumenta rápidamente en proporción a la eslora.*

Para llevar a cabo el entrenamiento del sistema, compilamos un corpus a partir de la rama catalana de Wikicorpus (Reese et al., 2010), siguiendo el mismo método que para el corpus de evaluación. Para cada término y su correspondiente artículo se extrae una oración definitoria (la primera). Para obtener las no definitorias se extraen aquellas oraciones en las que el término también aparece, con el fin de introducir un contexto en el que el número de distractores sea elevado e incrementar así la dificultad de la tarea.

3.2 Diseño Experimental

El preprocesado de los corpus se realiza con el etiquetador morfológico presente en Freeling (Carreras et al., 2004). Dado que vamos a explotar el potencial de Conditional Random Fields para etiquetado secuencial, los rasgos propuestos se aplican a nivel de token, y no a nivel de oración.

Partimos de una oración $s = f_1, f_2, ...f_n$, en la que cada f_i es un vector de rasgos que se corresponde con una palabra, y que recibe una etiqueta BIO dependiendo de si se encuentra al principio (Beginning), dentro (Inside) o fuera (Outside) de una definición. Este esquema de etiquetado permitirá, a posteriori, evaluar el rendimiento del algoritmo para cada etiqueta, siendo la etiqueta "B" un elemento clave en el ámbito de la detección de definiciones, dado que la primera palabra de una frase que contiene una definición es probable que sea (parte del) definiendum. A continuación, se describen los rasgos utilizados durante la fase de entrenamiento.

- **Surface**: La forma superficial de la palabra, tal y como aparece originalmente en el texto.
- **Lemma**: Forma lematizada de la palabra.
- **PoS**: Categoría gramatical.
- **Pos_Prob**: Probabilidad asignada por Freeling a la categoría gramatical de cada palabra.
- **BIO_NP**: En primer lugar, se aplica un filtro lingüístico para identificar sintagmas nominales. A continuación, se asignan etiquetas BIO a dichos sintagmas. Así, una oración quedaría etiquetada de la siguiente manera:
  - El[B-NP] verd[I-NP] és[O-NP] un[O-NP] dels[O-NP] tres[O-NP] colors[B-NP] primaris[I-NP] additius[I-NP] .[O-NP]

- *El[B-NP] verde[I-NP] es[O-NP] uno[O-NP] de[O-NP] los[O-NP] tres[O-NP] colores[B-NP] primarios[I-NP] aditivos[I-NP] .[O-NP]*

- **Def-TF**: La frecuencia de la palabra en las definiciones del corpus de entrenamiento.
- **Gen-TF**: La frecuencia de la palabra en un corpus de ámbito general, extraído del subcorpus catalán de *HC Corpora*[2], formado por documentos del género periodístico. Partimos de la hipótesis de que ciertas palabras o expresiones utilizados habitualmente en definiciones pueden ser poco frecuentes en texto del dominio general, como "es considera" (*se considera*) o "es defineix com" (*se define como*).
- **Def-TFIDF**: Se computa la métrica *Term Frequency - Inverse Document Frequency* para la palabra, considerando su frecuencia en las definiciones del corpus de entrenamiento, y tomando cada oración como documento. Esta métrica se define como:

$$\text{tfidf}(w, d, D) = \text{tf}(w, d) \times \text{idf}(w, D)$$

donde $\text{tf}(w, d)$ es la frecuencia de la palabra w en el documento d. Asimismo, $\text{idf}(w, D)$ se define como

$$\frac{|D|}{|\{d \in D : w \in d\}|}$$

donde D es el la colección de documentos y $|D|$ su cardinalidad.

- **Gen-TFIDF**: Tomando el mismo enfoque que en el rasgo anterior, computamos esta métrica para cada palabra tomando como referencia el corpus mencionado en el rasgo **Gen-TF**.
- **Termhood**: Esta métrica determina el grado de importancia de un candidato unipalabra a término en un dominio concreto (Kit y Liu, 2008), midiendo su frecuencia en un corpus general y un corpus específico. Se obtiene a través de la siguiente fórmula:

$$\text{Termhood(w)} = \frac{\text{r}_\text{D}(\text{w})}{|\text{V}_\text{D}|} - \frac{\text{r}_\text{B}(\text{w})}{|\text{V}_\text{B}|}$$

Donde r_D es el ránking por frecuencia de la palabra w en un corpus específico (en nuestro, caso, el corpus de entrenamiento), y r_B es el ránking por frecuencia de dicha palabra en el corpus general. Los denominadores se refieren al tamaño de cada corpus.

- **BIO_D y BIO_d**: En cada oración (definición o no) del corpus de entrenamiento, tomamos el primer verbo y asignamos la etiqueta *definiendum* (D) a lo que le precede. Por su parte, las palabras que le siguen reciben la etiqueta *definiens* (d). A continuación, identificamos los sintagmas nominales siguiendo el mismo procedimiento que en el rasgo BIO_NP. Así, por ejemplo, la oración anterior quedaría etiquetada:

El[B-definiendum] verd[i-definiendum] és[O-definiens] un[O-definiens] dels[O-definiens] tres[O-definiens] colors[B-definiens] primaris[I-definiens] additius[I-definiens] .[O-definiens]

- **Definitional prominence**: Introducimos la noción de prominencia definitoria, con el objetivo de establecer la probabilidad de una palabra w de aparecer en una oración definitoria ($s = \text{def}$). Para ello, consideramos su frecuencia en definiciones y oraciones no definitorias del corpus de entrenamiento en la siguiente ecuación:

$$\text{DefProm(w)} = \frac{\text{DF}}{|\text{Defs}|} - \frac{\text{NF}}{|\text{Nodefs}|}$$

donde $\text{DF} = \sum_{i=0}^{i=n}(s_i = \text{def} \wedge w \in s_i)$ y $\text{NF} = \sum_{i=0}^{i=n}(s_i = \text{nodef} \wedge w \in s_i)$.

- **Definiendum prominence**: Partiendo de la hipótesis de que el hecho de que una palabra aparezca frecuentemente en posición de posible *definiendum* puede ser un indicador de su papel en definiciones, este rasgo viene dado por

$$\text{DP(w)} = \frac{\sum_{i=0}^{i=n} w_i \in \text{term}_D}{|DT|}$$

[2] http://www.corpora.heliohost.org/

donde $term_D$ es un sintagma nominal (es decir, un candidato a término) que aparece en posición de definiendum. Finalmente, |DT| se refiere al tamaño del corpus de terminología de definienda.

- **Definiens prominence**: Este rasgo consiste en la misma ecuación que en el caso anterior, esta vez considerando términos que aparecen en posición de posible definiens.

El algoritmo CRF itera sobre cada uno de los vectores y aprende combinaciones de los rasgos descritos. Estas combinaciones se establecen de antemano, por ejemplo, para aprender como rasgo la combinación lema + termhood de la palabra anterior y la combinación lema + categoría gramatical + definitional_prominence de la palabra actual. La Sección 4 describe los resultados obtenidos tras llevar a cabo experimentos realizados con varias configuraciones de rasgos.

4 Evaluación

Se han realizado experimentos que combinan los rasgos descritos en la sección 3.2, así como su combinatoria. Ofrecemos resultados en términos de Precisión, Cobertura, y F-Measure para cada una de las clases consideradas (B, O, I), además de una media de las 3. Éstas se aplican a nivel de palabra, y de la forma más restrictiva posible. Es decir, que se considera un error cuando el algoritmo predice correctamente que una palabra se encuentra en una definición, pero asigna una categoría incorrecta (es decir, Beginning en vez de Inside o viceversa). Con estas consideraciones, realizamos cuatro configuraciones experimentales, a saber:

- **Baseline**: Esta configuración sólo considera la forma superficial del token en la iteración actual.
- **C-1**: Se toman en cuenta rasgos lingüísticos (forma superficial, lema, categoría gramatical y pertenencia a sintagma nominal) sobre una ventana [i-3:i+3], siendo i la posición de la iteración actual.
- **C-2**: Se toman únicamente rasgos estadísticos (tf-def, tf-gen, tfidf-def, tfidf-gen, termhood, definitional prominence, definiendum prominence y definiens prominence). La ventana es la misma que para **C-1**.
- **C-3**: Toma en cuenta todos los rasgos.

La Tabla 1 muestra los resultados obtenidos con estas cuatro configuraciones. Los identificadores de las filas se refieren a: (1) Si el resultado es en precisión (P), cobertura (C), o F-Measure (F), y (2) si la evaluación corresponde a la etiqueta Beginning (B), Inside (I), Outside (O), o a la media de las tres (M), que en definitiva refleja el comportamiento general del sistema propuesto en este artículo. Se puede observar que a partir de una baseline que obtiene un 67.31 de F-Measure, ésta es superada por la combinación de rasgos lingüísticos (en C-1), que obtiene un F=75.85, y rasgos estadísticos (C-2), que llega a F=75.68. La combinación de ambos conjuntos de rasgos obtiene resultados altamente competitivos (F=86.69), lo cual sugiere que ambos conjuntos de rasgos son informativos y contribuyen al proceso de aprendizaje.

	Baseline	C-1	C-2	C-3
P-B	67.50	89.29	80.68	**93.60**
C-B	51.72	57.47	**88.62**	85.87
F-B	58.57	69.93	84.47	**89.57**
P-I	58.49	84.89	72.25	**90.71**
C-I	49.58	51.82	**88.80**	83.48
F-I	53.67	64.35	79.68	**86.95**
P-O	88.03	**89.19**	77.24	79.36
C-O	91.43	**97.76**	53.08	88.23
F-O	89.79	**93.28**	62.92	83.56
P-M	71.34	87.78	76.72	**87.89**
C-M	64.24	69.01	76.83	**85.85**
F-M	67.31	75.85	75.68	**86.69**

Tabla 1: Resultados obtenidos en término de precisión, cobertura y F-measure

4.1 Discusión

A la luz de los resultados obtenidos, cabe destacar el importante papel que juegan los rasgos lingüísticos (categoría gramatical, lema y pertenencia o no a un sintagma nominal), ya que observamos un mejor rendimiento de un modelo entrenado sólo con rasgos de este tipo, en comparación con un sistema entrenado sólo con rasgos estadísticos. No obstante, la combinación de rasgos de ambos tipos contribuye al mayor rendimiento de las tres configuraciones propuestas.

Observando algunos de las instancias incorrectamente clasificadas, se observa una tendencia al sobre-entrenamiento con respecto a las definiciones basadas en el modelo *genus et differentia* propio de Viquipèdia. Por ejemplo, la definición del término "gas natural" es:

- El gas natural és una font d'energia fòssil que, com el carbó o el petroli, està constituïda per una barreja d'hidrocarburs, unes molècules formades per àtoms de carboni i hidrogen.
- *El gas natural es una fuente de energía fósil que, como el carbón o el petróleo, está constituída por una mezcla de hidrocarburos, unas moléculas formadas por átomos de carbono e hidrógeno.*

Por su parte, en nuestro test set, además de la oración anterior, existe el siguiente distractor:

- **El gas natural és una** energia primària, o que es pot obtenir directament sense **transformació**
- ***El gas natural es una*** *energía primaria, o que se puede obtener directamente sin* ***transformación***

Las palabras resaltadas en negrita fueron incorrectamente marcadas como pertenecientes a una definición por nuestro clasificador. Dos conclusiones se pueden extraer de casos como éste: (1) Existen relaciones semánticas entre conceptos que podrían ser consideradas como definitorias, según un criterio ligeramente más laxo, y esto se refleja en algunos de los falsos negativos obtenidos en nuestra evaluación; (2) Asumimos que en el futuro sería deseable contar con una heurística post-clasificación (o un segundo proceso de clasificación) para realizar un segundo proceso de clasificación sobre palabras que o bien han sido clasificadas con poca probabilidad, o bien tienen palabras próximas clasificadas de otra manera. Por ejemplo, en una oración en la que el 80 % de las palabras han sido clasificadas como no definitorias, es razonable asumir que si el 20 % fueron clasificadas como definitorias, la probabilidad de que ésta sea una clasificación incorrecta es elevada.

Finalmente, con respecto a los rasgos estadísticos, su introducción al proceso de entrenamiento, si bien contribuyen a mejorar el sistema, no parece que puedan constituir la base de un sistema de extracción de definiciones, o al menos, no sería recomendable descartar rasgos estadísticos. A continuación, se describen algunas de las posibles razones por las que los rasgos estadísticos propuestos son susceptibles de mejora:

- La falta de un paso previo en identificación de terminología para generalizar los términos en posición de definiendum. Esto provoca que sólo aquellos términos multipalabra con alguna palabra repetida se benefician de métricas que toman en cuenta la frecuencia de sus componentes. Éste es el caso, por ejemplo, de nombres propios que comparten algún apellido o definienda que comparten algún término (por ejemplo, en especies de peces como: **ammodytes** tobianus, **ammodytes** immaculatus o **ammodytes** marinus, entre otros).
- Posible falta de representatividad de los corpus de referencia. Si bien el dominio en el que se ha desarrollado este estudio es homogéneo en el género textual (enciclopédico), podemos afirmar que se trata de un estudio no delimitado a un dominio concreto. Nuestra hipótesis es que métricas como *tfidf* o *definitional_prominence* serían más informativas aplicadas a dominios concretos. De hecho, en el campo de ED, salvo contadas excepciones (Snow, Jurafsky y Ng, 2004; Velardi, Navigli y D'Amadio, 2008; Cui, Kan y Chua, 2005), la tendencia es desarrollar y evaluar sistemas en corpora pertenecientes a un dominio específico.

5 Conclusiones

En este trabajo se ha descrito un sistema de extracción de definiciones para el catalán. Los corpora de entrenamiento y test, similar a los datasets descritos en Navigli, Velardi y Ruiz-Martínez (2010), son obtenidos de Viquipèdia. El corpus de entrenamiento es un subcorpus de la rama catalan de Wikicorpus (Reese et al., 2010), mientras que el corpus de test ha sido validado manualmente. Afrontamos el problema como una tarea de clasificación secuencial supervisada, en la que a cada palabra se le asigna la etiqueta BIO, dependiendo de si el sistema predice que se encuentra al principio (Beginning), dentro (Inside) o fuera (Outside) de una definición.

Utilizamos el algoritmo Conditional Random Fields, y combinamos rasgos lingüísticos y estadísticos, obteniendo resultados razonablemente prometedores.

6 Trabajo Futuro

Tras haber explorado el proceso de desarrollar un sistema de ED para el catalán, a continuación se enumeran algunos aspectos susceptibles de mejora, además de posibles líneas de trabajo futuro: (1) un estudio de la relevancia de los distintos rasgos y su combinatoria ayudaría a desarrollar un sistema más eficiente y que no considerara rasgos escasamente discriminatorios. (2) Realizar experimentos con distintos corpus de referencia, y con distintos ratios entre definiciones y no definiciones ayudaría a valorar la influencia de estos datasets en el proceso de aprendizaje. Finalmente, (3) aplicar distintos algoritmos de ED que existen para el inglés daría una idea del comportamiento de nuestro sistema en un contexto comparativo.

También creemos que el proceso de evaluación (en este trabajo y en ED en general) se puede desarrollar y especificar más si: (1) se realiza una evaluación de contenido. Este enfoque nos permitiría determinar si se han dado casos de definiciones extraídas parcialmente. Y (2), si se realizara una evaluación en distintas etapas del flujo (*pipeline*) del sistema, como por ejemplo, para la identificación de definienda y definiens.

Agradecimientos

Agradecemos a los revisores anónimos sus comentarios y sugerencias. Este trabajo ha sido parcialmente financiado por el proyecto número TIN2012-38584-C06-03 del Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, España.

Bibliografía

Alarcón, Rodrigo. 2009. *Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios*. Ph.D. tesis, Universitat Pompeu Fabra.

Auger, Alain y Pierre Knecht. 1997. Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles.

Bontas, Elena Paslaru y Malgorzata Mochol. 2005. Towards a cost estimation model for ontology engineering. En Rainer Eckstein y Robert Tolksdorf, editores, *Berliner XML Tage*, páginas 153–160.

Carreras, Xavier, Isaac Chao, Lluis Padró, y Muntsa Padró. 2004. Freeling: An open-source suite of language analyzers. En *LREC*.

Cui, Hang, Min-Yen Kan, y Tat-Seng Chua. 2005. Generic soft pattern models for definitional question answering. En *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, páginas 384–391. ACM.

Del Gaudio, Rosa, Gustavo Batista, y António Branco. 2013. Coping with highly imbalanced datasets: A case study with definition extraction in a multilingual setting. *Natural Language Engineering*, páginas 1–33.

Espinosa-Anke, Luis. 2013. Towards definition extraction using conditional random fields. En *Proceedings of RANLP 2013 Student Research Workshop*, páginas 63–70.

Feliu, Judit y M Teresa Cabré. 2002. Conceptual relations in specialized texts: new typology and an extraction system proposal. En *TKE 02, 6th International Conference in Terminology and Knowledge Engineering*, páginas 45–49.

Kit, Chunyu y Xiaoyue Liu. 2008. Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*, 14(2).

Lafferty, John D., Andrew McCallum, y Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. En *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, páginas 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Meyer, Ingrid. 2001. Extracting knowledge-rich contexts for terminography. *Recent advances in computational terminology*, 2:279.

Muresan, A y Judith Klavans. 2002. A method for automatically building and evaluating dictionary resources. En *Proceedings*

of the Language Resources and Evaluation Conference (LREC.

Nakamura, Jun-ichi y Makoto Nagao. 1988. Extraction of semantic information from an ordinary english dictionary and its evaluation. En *Proceedings of the 12th Conference on Computational Linguistics - Volume 2*, COLING '88, páginas 459–464, Stroudsburg, PA, USA. Association for Computational Linguistics.

Navigli, Roberto y Paola Velardi. 2010. Learning word-class lattices for definition and hypernym extraction. En *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, páginas 1318–1327, Stroudsburg, PA, USA. Association for Computational Linguistics.

Navigli, Roberto, Paola Velardi, y Juana María Ruiz-Martínez. 2010. An annotated dataset for extracting definitions and hypernyms from the web. En *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, páginas 3716–3722, Valletta, Malta. European Language Resources Association (ELRA).

Okazaki, Naoaki. 2007. Crfsuite: a fast implementation of conditional random fields (CRFs).

Park, Youngja, Roy J. Byrd, y Branimir K. Boguraev. 2002. Automatic Glossary Extraction: Beyond Terminology Identification. En *Proceedings of the 19th International Conference on Computational Linguistics*, páginas 1–7. Association for Computational Linguistics.

Reese, Samuel, Gemma Boleda, Montse Cuadros, Lluís Padró, y German Rigau. 2010. Wikicorpus: A word-sense disambiguated multilingual wikipedia corpus. En *LREC*. European Language Resources Association.

Saggion, Horacio y Robert Gaizauskas. 2004. Mining on-line sources for definition knowledge. En *17th FLAIRS*, Miami Bearch, Florida.

Sierra, Gerardo, Rodrigo Alarcón, César Aguilar, y Alberto Barrón. 2006. Towards the building of a corpus of definitional contexts. En *Proceeding of the 12th EURALEX International Congress, Torino, Italy*, páginas 229–40.

Snow, Rion, Daniel Jurafsky, y Andrew Y Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

Trimble, L. 1985. *English for Science and Technology: A Discourse Approach*. Cambridge Language Teaching Library.

Velardi, Paola, Roberto Navigli, y Pierluigi D'Amadio. 2008. Mining the web to create specialized glossaries. *IEEE Intelligent Systems*, 23(5):18–25, Septiembre.

Westerhout, Eline y Paola Monachesi. 2007. Extraction of Dutch definitory contexts for elearning purposes. *Proceedings of the Computational Linguistics in the Netherlands (CLIN 2007), Nijmegen, Netherlands*, páginas 219–34.