Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 131-134 recibido 15-03-2018 revisado 28-04-2018 aceptado 03-05-2018

Extracción automática de equivalentes multilingües de colocaciones

Automatic extraction of multilingual collocation equivalents

Marcos Garcia
Universidade da Coruña
Grupo LyS, Departamento de Letras, Facultade de Filoloxía
Campus da Zapateira, 15008, A Coruña, Galiza
marcos.garcia.gonzalez@udc.gal

Resumen: Este trabajo presenta el proyecto *Extracción automática de equivalentes multilingües de colocaciones*, financiado por el programa *Becas Leonardo a Investigadores y Creadores Culturales 2017* dentro del área de Humanidades. El objetivo principal del proyecto es extraer automáticamente equivalentes multilingües de colocaciones fraseológicas. Para ello proponemos diversas estrategias que combinan análisis sintáctico y estadístico con técnicas de semántica distribucional, tanto monolingües como multilingües. Las lenguas de trabajo del proyecto son el portugués, el español y el inglés.
Palabras clave: colocaciones, fraseología, léxico, sintaxis, semántica distribucional

Abstract: This work presents the project *Automatic extraction of multilingual collocation equivalents*, funded by a *2017 Leonardo Grant for Researchers and Cultural Creators*, in the area of Humanities. The main objective of this project is to automatically extract multilingual equivalents of phraseological collocations. We propose several strategies which combine syntactic and statistical analysis with distributional semantic techniques, both monolingual and multilingual. The languages of the project are Portuguese, Spanish, and English.
Keywords: collocations, phraseology, lexicon, syntax, distributional semantics

1 Introducción y objetivos

Un alto porcentaje de las expresiones lingüísticas que producimos los hablantes de una lengua está formado por estructuras más o menos preconstruidas de elementos léxicos (Erman y Warren, 2000). Entre estas estructuras encontramos diferentes tipos, tales como las locuciones (expresiones no composicionales desde un punto de vista semántico), o las colocaciones (entendidas aquí como combinaciones composicionales restringidas léxicamente) (Mel'čuk, 1995; Alonso-Ramos, 1995).

A pesar de que estas expresiones no presentan grandes problemas en el proceso de adquisición de las lenguas maternas, sí que generan dificultades en situaciones en las que coexiste más de un idioma, como en el aprendizaje de lenguas extranjeras (Altenberg y Granger, 2001), o en sistemas de traducción automática (Orliac y Dillinger, 2003).

Así, la colocación del inglés *brown sugar*, podría tener como equivalentes *azúcar moreno* –en español–, o *açúcar mascavado* –en portugués–, pero no otras alternativas como **azúcar marrón*, **açúcar castanho*, etc.

Conocer qué combinaciones son posibles y cuáles no son habituales en un dado idioma puede ser útil para diversos propósitos, desde la enseñanza de lenguas o la traducción automática, a la compresión y generación de lenguaje natural.

Teniendo esto en cuenta, el objetivo principal de este proyecto consiste en implementar métodos basados en lingüística computacional que permitan extraer automáticamente, con alta precisión y a gran escala equivalentes de colocaciones en portugués, español e inglés.

Especial interés tendrá la obtención de equivalentes *incongruentes* de colocaciones, i.e., aquellos en los que la traducción de sus constituyentes no es coherente (Nesselhauf, 2003). Por ejemplo, el par inglés-español "pay attention"–"prestar atención", a diferencia de equivalentes *congruentes* como "formulate

ISSN 1135-5948. DOI 10.26342/2018-61-15

[a] hypothesis"–"formular [una] hipótesis".

En el proceso de obtención de equivalentes multilingües de colocaciones podemos identificar dos tareas bien diferenciadas:

1. Extracción de colocaciones monolingües.
2. Identificación de equivalentes multilingües.

Para la primera utilizaremos análisis sintáctico de dependencias y medidas de asociación estadística. Además, compararemos métodos distribucionales composicionales y no composicionales con el objetivo de discriminar aquellas colocaciones fraseológicas de las puramente estadísticas.

Para identificar equivalentes de colocaciones en diferentes idiomas usaremos modelos multilingües de semántica distribucional, que serán obtenidos tanto de corpus paralelos y comparables como de textos monolingües.

Este proyecto continúa y amplía un conjunto de trabajos previos sobre la extracción automática de colocaciones multilingües (Garcia, García-Salido, y Alonso-Ramos, 2018; Garcia, García-Salido, y Alonso-Ramos, 2017) y, del mismo modo, publicará y distribuirá utilizando licencias libre todos los recursos generados durante su realización.

2 Metodología y plan de trabajo

Para llevar a cabo nuestro objetivo, el proyecto se divide en cuatro partes, cada una de ellas organizada en diferentes subtareas:

1. Compilación, análisis y clasificación de corpus paralelos, comparables y monolingües.
2. Identificación de candidatos a colocaciones mediante técnicas lingüístico-estadísticas.
3. Creación y evaluación de modelos multilingües de semántica distribucional.
4. Clasificación automática de equivalentes multilingües de colocaciones.

La primera etapa (compilación y análisis de corpus) consistirá en obtener a través de la web grandes cantidades de corpus, tanto multilingües (paralelos y comparables) como monolingües. Los diferentes recursos compilados pertenecerán a diferentes registros y dominios (enciclopédico, periodístico, oral, etc.), con el fin de conseguir un conjunto amplio y diverso de información fraseológica y distribucional. Los corpus serán clasificados, utilizando técnicas semiautomáticas, en función del registro y de la variedad lingüística.

En la segunda fase aplicaremos métodos híbridos, que combinan información sintáctica y fraseológica con modelos estadísticos, para la identificación de candidatos a colocaciones monolingües (Seretan, 2011; Evert et al., 2017). Aquí será necesario realizar un análisis automático de los corpus utilizando sistemas de procesamiento del lenguaje natural (Garcia y Gamallo, 2015; Straka y Straková, 2017). El análisis sintáctico será realizado con *Universal Dependencies*, lo que nos permitirá, entre otras cosas, obtener corpus etiquetados con una anotación homogénea en las diversas lenguas de trabajo (Nivre, 2015). Además, evaluaremos diferentes estrategias composicionales y no composicionales para diferenciar colocaciones fraseológicas de combinaciones estadísticas (Pearce, 2001; Kiela y Clark, 2013; Rodríguez-Fernández et al., 2016; Farahmand y Henderson, 2016).

El resultado de esta segunda fase serán grandes listas monolingües de colocaciones candidatas, ordenadas tanto mediante información estadística (*log-likelihood*, *DeltaP*, etc.), como por su carácter fraseológico.

Denominamos colocaciones fraseológicas a aquellas combinaciones de dos unidades léxicas (*base* y *colocativo*) en las que el significado de la base se mantiene intacto en la colocación, mientras que el del colocativo depende del significado de la propia colocación (e.g., $odio_{Base}$ $mortal_{Colocativo}$) (Mel'čuk, 1995). A diferencia de las visiones puramente estadísticas de las colocaciones, la tradición fraseológica exige una relación sintáctica directa entre los dos elementos de una colocación. Además, la elección del colocativo no es libre, sino que está restringida por la base (Alonso-Ramos, 1995).

En este proyecto nos centraremos en tres tipos diferentes de colocaciones:

- Verbo-Objeto: $ceño_b$, $fruncir_c$ ("fruncía el ceño"); $moción_b$, $secundar_c$ ("secundaron la moción").
- Nombre-Adjetivo: $dinero_b$, $negro_c$; $saludo_b$, $cordial_c$.
- Nombre-Nombre: $coral_b$, $arrecife_c$ ("arrecifes de coral"); $lana_b$, $ovillo_c$ ("ovillo de lana").

Téngase en cuenta que el uso de dependencias sintácticas permite identificar colocaciones a larga distancia ("fruncía con dolor pero rápidamente el ceño") y en diferente orden ("saludos$_{\text{nombre}}$ cordiales$_{\text{adjetivo}}$", "mayor$_{\text{adjetivo}}$ cantidad$_{\text{nombre}}$"). Además, el análisis en dependencias universales establece relaciones sintácticas entre palabras léxicas (e.g., *cartón*→`nmod`→*tabaco* en "cartón de tabaco"), lo que agiliza el proceso de extracción de candidatos a colocaciones.

Tanto la clasificación de los corpus multilingües realizada en la primera etapa, como la información obtenida mediante el análisis computacional, serán utilizadas en la tercera parte del proyecto, cuyo objetivo es la creación de modelos de semántica distribucional multilingües.

Así, en esta tercera fase crearemos diferentes modelos bilingües (portugués-inglés, portugués-español y español-inglés) utilizando corpus paralelos, comparables y monolingües. Además, cada par de idiomas contará en esta etapa con modelos construidos utilizando tres estrategias: (i) monoléxica, en los que cada vector representa una única palabra (ortográfica), (ii) contextual, en los que los diferentes sentidos de cada palabra serán agrupados en función de su distribución semántica, y (iii) no composicional, que representa en un único vector las dos unidades léxicas de cada colocación candidata.

Durante esta fase evaluaremos la precisión de los diferentes modelos distribucionales en la identificación de equivalentes multilingües de dos tipos de colocaciones: (i) *congruentes*, en las que la traducción de las unidades léxicas es coherente interlingüísticamente (por ejemplo, entre inglés y español: *vermouth$_b$, red$_c$ – vermú$_b$, rojo$_c$*), e (ii) *incongruentes*, en las que el carácter impredecible de las colocaciones implica que los equivalentes multilingües no sean traducciones directas de la base y el colocativo (*wine$_b$, red$_c$ – vino$_b$, tinto$_c$*).

En la última parte del proyecto utilizaremos los candidatos a colocaciones monolingües (obtenidos en la fase 2) y los modelos distribucionales (fase 3) para generar los equivalentes multilingües de colocaciones. Primero, unificaremos y seleccionaremos las colocaciones candidatas obtenidas por las diferentes medidas de asociación estadística y de análisis fraseológico, generando así listas de colocaciones de alta confianza para cada uno de los tipos (verbo-objeto, adjetivo-nombre y nombre-nombre). Después, utilizando estas listas monolingües de colocaciones, aplicaremos los modelos distribucionales para identificar (i) equivalentes de la base y del colocativo mediante similaridad semántica (modelos monoléxicos, aplicados principalmente a equivalentes congruentes), y (ii) equivalentes de la colocación como combinación contextual o como unidad (modelos contextuales y no composicionales, para la identificación de equivalentes incongruentes).

Por último, los métodos de mayor precisión y cobertura se aplicarán en los tres pares de lenguas analizadas (portugués-inglés, portugués-español y español-inglés), extrayendo así grandes cantidades de equivalentes bilingües de colocaciones. Una vez obtenidas estas listas bilingües, se aplicará un método de fusión por transitividad para generar un recurso multilingüe portugués-inglés-español. Este proceso permitirá también aumentar el número de equivalentes bilingües, ya que podrán descubrirse, por ejemplo, pares portugués-inglés no reconocidos previamente, a través de equivalentes español-portugués y español-inglés correctamente identificados.

3 Equipo de trabajo

El proyecto se lleva a cabo en el Grupo LyS (Lengua y Sociedad de la Información), de la Universidade da Coruña. El Grupo LyS es un equipo interdisciplinar de investigación en Lingüística Computacional del que forman parte diferentes profesores, investigadores y estudiantes tanto del área de Lingüística General como de Ciencias de la Computación.

Además del investigador principal del proyecto, en diferentes etapas del mismo serán contratadas dos personas que colaboren tanto en las tareas de compilación, anotación y análisis de los corpus, como en el diseño, implementación y evaluación de las varias estrategias que analizaremos.

Por otro lado, el hecho de formar parte de un grupo interdisciplinar nos permite trabajar en estrecha colaboración con otros miembros del equipo que, sin ser parte directa del proyecto, aportan conocimientos fundamentales para su desarrollo.

Agradecimientos

Proyecto realizado con una Beca Leonardo a Investigadores y Creadores Culturales 2017

(Fundación BBVA), y parcialmente financiado por un contrato *Juan de la Cierva - incorporación* (IJCI-2016-29598, Ministerio de Economía y Competitividad) y por *RELEX: Rede de Lexicografía* (ED341D R2016/046).

Bibliografía

Alonso-Ramos, M. 1995. Hacia una definición del concepto de colocación: de J.R. Firth a I.A. Mel'čuk. *Revista de Lexicografía*, 1:9–28.

Altenberg, B. y S. Granger. 2001. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, 22(2):173–195.

Erman, B. y B. Warren. 2000. The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse*, 20(1):29–62.

Evert, S., P. Uhrig, S. Bartsch, y T. Proisl. 2017. E-VIEW-alation–a Large-scale Evaluation Study of Association Measures for Collocation Identification. En *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, páginas 531–549, Leiden.

Farahmand, M. y J. Henderson. 2016. Modeling the non-substitutability of multiword expressions with distributional semantics and a log-linear model. En *Proceedings of the 12th Workshop on Multiword Expressions (MWE 2016) at ACL 2016*, páginas 61–66, Berlin. ACL.

Garcia, M. y P. Gamallo. 2015. Yet another suite of multilingual NLP tools. En José-Luis Sierra-Rodríguez and José Paulo Leal and Alberto Simões, editor, *Languages, Applications and Technologies. Communications in Computer and Information Science*, International Symposium on Languages, Applications and Technologies (SLATE 2015), páginas 65–75. Springer.

Garcia, M., M. García-Salido, y M. Alonso-Ramos. 2017. Using bilingual word-embeddings for multilingual collocation extraction. En *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017) at EACL 2017*, páginas 21–30, Valencia. ACL.

Garcia, M., M. García-Salido, y M. Alonso-Ramos. 2018. Discovering bilingual collocations in parallel corpora: A first attempt at using distributional semantics. En I. Doval y M. T. Sánchez Nieto, editores, *Parallel corpora for contrastive and translation studies: New resources and applications*. John Benjamins Publishing.

Kiela, D. y S. Clark. 2013. Detecting compositionality of multi-word expressions using nearest neighbours in vector space models. En *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, páginas 1427–1432, Copenhagen.

Mel'čuk, I. 1995. Phrasemes in language and phraseology in linguistics. En *Idioms: Structural and psychological perspectives*. Lawrence Erlbaum Associates, páginas 167–232.

Nesselhauf, N. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied linguistics*, 24(2):223–242.

Nivre, J. 2015. Towards a Universal Grammar for Natural Language Processing. En *International Conference on Intelligent Text Processing and Computational Linguistics*, páginas 3–16. Springer.

Orliac, B. y M. Dillinger. 2003. Collocation extraction for machine translation. En *Proceedings of Ninth Machine Translation Summit (MT Summit IX)*, páginas 292–298, New Orleans, Lousiana.

Pearce, D. 2001. Synonymy in collocation extraction. En *Proceedings of the Workshop on WordNet and other lexical resources at NAACL 2001*, páginas 41–46, Pittsburgh. ACL.

Rodríguez-Fernández, S., L. Espinosa Anke, R. Carlini, y L. Wanner. 2016. Semantics-driven recognition of collocations using word embeddings. En *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, páginas 499–505, Berlin.

Seretan, V. 2011. *Syntax-based collocation extraction*, volumen 44 de *Text, Speech and Language Technology Series*. Springer Science & Business Media.

Straka, M. y J. Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. En *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, páginas 88–99, Vancouver.