Predicción de expresiones fijas en textos

Fernando Sánchez Vega[1], Luis Villaseñor Pineda[1],
Luis Meneses Lerín[2], Salah Mejri[2]

[1] Laboratorio de Tecnologías del Lenguaje,
Coordinación de Ciencias Computacionales,
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), México.

[2] LDI, (Lexiques, Dictionnaires, Informatique)
UMR 7187, CNRS/Université Paris 13 (Sorbonne Paris Cité) et
Université de Cergy Pontoise, France

Reporte Técnico No. CCC-16-004
Marzo de 2016

Luis Enrique Erro 1
Sta. Ma. Tonantzintla,
72840, Puebla, México.

CONTENIDO

1 Introducción 5

2 Predicción de expresión fijas en textos 6

2.1 Recuperación de expresiones fijas candidatas 6

2.1.1 Medida de opacidad semántica 7

2.1.2 Experimentos 8

2.2 Detección de expresiones fijas 10

2.2.1 Experimentos 10

3 Predicción de intención comunicativa y polaridad 11

3.1 Experimentos 12

4 Conclusiones 14

5 Referencias 14

1 INTRODUCCIÓN

En este documento se presentan las actividades desarrolladas durante las estancias de investigación en el contexto del proyecto *Diccionarios electrónicos monolingües coordinados de expresiones fijas francés-español (España)-español (México)* del Programa ECOS-ANUIES-SEP, CONACYT (No. Ref. ECOS M11-H04). 2012-2015.

El proyecto que enmarca los trabajos aquí presentados tiene como principal objetivo la generación de recursos y técnicas especializadas en la obtención de diccionarios electrónicos de expresiones fijas mediante la colaboración de grupos multidisciplinarios especializados en la lingüística fraseológica y la lingüística computacional.

Los diccionarios electrónicos monolingües coordinados constituyen uno de los más importantes productos de la lexicografía contemporánea (Aguila Escobar, 2006). La coordinación de estos diccionarios permite tener vínculos que relacionan las identidades contenidas en diccionarios de diferentes lenguas y/o variantes de una lengua. Por sus características, estos diccionarios son de gran utilidad para estudios de lexicografía comparada entre lenguas y constituyen herramientas ampliamente utilizadas en la traducciones de textos (Papadopoulou, 2010).

Una de las clases más interesantes de identidades incluidas en los diccionarios monolingües coordinados son las expresiones fijas. Las expresiones fijas son expresiones poli-lexicales cuyo significado no está determinado de forma composicional, es decir, el sentido de la expresión no es constituido por el aporte de cada uno de los elementos (Gross, 1996), condición de no composicionalidad (u opacidad semántica). El sentido de las expresiones fijas implica que se deben entender y tratar como una unidad léxica y no como palabras aisladas. En la traducción de un texto hay que tener especial cuidado con las expresiones fijas, por ello es de gran utilidad la existencia de diccionarios monolingües coordinados especializados en expresiones fijas.

Los diccionarios especializados en expresiones fijas son muy útiles, pero su elaboración implica múltiples desafíos, pues no es sencillo identificar las expresiones fijas de forma masiva y automática. Por ello la creación de estos diccionarios requiere de herramientas automáticas y procesos semiautomáticos que deben ser validados manualmente durante diferentes etapas de la creación de los diccionarios.

Los trabajos aquí presentados abordan dos importantes aspectos de las expresiones fijas que han aportado herramientas útiles en la construcción de diccionarios monolingües coordinados de

expresiones fijas. El primer aspecto abordado es la opacidad semántica, que ha sido empleada para identificar una expresión fija en un texto de forma automática mediante métodos computacionales. El segundo aspecto involucra la observación de las expresiones fijas en la expresión informal (y espontánea), así como su relación con la comunicación objetiva y subjetiva para la predicción de la intención comunicativa (objetiva y subjetiva) de las frases espontaneas mediante métodos computacionales.

2 PREDICCIÓN DE EXPRESIÓN FIJAS EN TEXTOS

La predicción de la fijeza de las palabras permite evaluar cuál de las palabras en un texto son parte de una expresión fija y, por tanto, deben ser entendidas y tratadas como una unidad de expresión poli-lexical.

En el trabajo presentado aquí, realizado en el entorno del proyecto ECOS-ANUIES-SEP, CONACYT, antes citado, se han desarrollado dos herramientas que apoyan la conformación de los diccionarios de expresiones fijas en dos etapas: 1) la recuperación de expresiones fijas candidatas, y 2) la predicción de la opacidad de la posible expresión fija.

2.1 RECUPERACIÓN DE EXPRESIONES FIJAS CANDIDATAS

El objetivo de la herramienta generada en esta etapa es la recuperación de la mayor cantidad de expresiones fijas que contengan un verbo ancla predefinido como parte de la expresión fija.

El método propuesto para la herramienta se basa en la hipótesis de la opacidad semántica de las expresiones fijas. El método evalúa automáticamente la opacidad de cada palabra en las frases que contienen el verbo ancla. Aquellas frases que contengan palabras con un alto grado de opacidad son recuperadas y evaluadas por los lingüistas fraseológicos para su posible incorporación a diccionarios de expresiones fijas.

Nuestra aproximación automática para la verificación de la opacidad de las palabas en una frase se realiza mediante la comparación de los campos semánticos de una palabra con el resto de palabras de la frase. Una distancia significativa entre el campo semántico de la palabra y el campo de la frase indica una alta probabilidad de que el uso de esta palabra sea metafórico y de que esté dentro de una expresión fija, en donde el sentido directo de la palabra no es cercano al sentido general del resto de la frase.

El proceso automático para la obtención del campo semántico se realiza mediante un modelo aproximado. El modelo computacional empleado para la captura del campo semántico de las palabras es el modelo del espacio vectorial (Stock y Stock, 2013). Este modelo computacional aproxima el modelo semántico de la palabra en un vector *n* dimensional $\vec{v}$ mediante la captura de las palabras con las que suele co-ocurrir en los textos. El modelo del espacio vectorial se basa en el supuesto de que el sentido de las palabras se puede deducir a partir del contexto en el que suelen emplearse éstas y, por tanto, la suma de los términos con los que co-ocurre puede capturar de forma aproximada un modelo de la semántica de la palabra.

El vector modelo semántico $\overrightarrow{w_i}$ de cada palabra w_i es un vector *n* dimensional donde *n* es determinado por el tamaño del vocabulario del corpus de donde se extrae el modelo semántico. El vector $\overrightarrow{w_i}$ contiene el número de veces que cada palabra del vocabulario co-ocurre con la palabra w_i, es decir en la dimensión *j* el vector $\overrightarrow{w_i}$ contendrá el número de veces que w_i y w_j co-ocurren en el corpus en donde se calcula el modelo semántico.

2.1.1 Medida de opacidad semántica

Dada una frase *F* de *m* palabras, que contienen el verbo ancla v_k, la evaluación de opacidad semántica O_s (ecuación 1) de la expresión fija E_f que contiene el verbo de interés v_k y la palabra w_i está determinada por la similitud coseno (Stock y Stock, 2013) del vector de la expresión fija $\overrightarrow{E_f}$ y el vector de su contexto $\overrightarrow{V_c}$ dado por el vector suma de las *m-2* palabras de la frase (se excluyen la palabra w_i y el verbo ancla v_k).

$$O_s\left(\overrightarrow{E_f} \cdot \overrightarrow{V_c}\right) = sim_{coseno}\left(\overrightarrow{E_f}, \overrightarrow{V_c}\right) = \frac{\overrightarrow{E_f} \cdot \overrightarrow{V_c}}{|\overrightarrow{E_f}||\overrightarrow{V_c}|} \quad (1)$$

dónde:

$$\overrightarrow{E_f} = \overrightarrow{w_i} + \overrightarrow{v_k} \quad (2)$$

$$\overrightarrow{V_c} = -\overrightarrow{E_f} + \sum_{\overrightarrow{w_x} \in \{F\}} \overrightarrow{w_x} \quad (3)$$

La medida de opacidad semántica permite calcular una probabilidad de que en la frase *F* exista una expresión fija que contiene el verbo v_k y la palabra w_i. Este valor nos permite ordenar las frases recuperadas presentando, primero, las que obtienen una mayor probabilidad de tener una frase fija.

2.1.2 Experimentos

Se realizaron múltiples experimentos de recuperación sobre el corpus de Periódicos Mexicanos (Meneses Lerin, 2013). El verbo ancla elegido para las pruebas es el verbo "dar", que ha sido ampliamente estudiado en este corpus (Meneses Lerin, 2013). Adicionalmente la condición de verbo de apoyo de "dar" resulta conveniente pues este tipo de verbos son frecuentemente encontrados en las expresiones fijas.

Comparamos el ordenamiento o *ranking* de las frases recuperadas empleando el método propuesto de la opacidad semántica con tres modificaciones. El objetivo de esta comparación es evaluar si nuestro método es la mejor forma de emplear los modelos computacionales de campos semánticos. Las tres modificaciones varían la composición de los vectores $\overrightarrow{E_f}$ y $\overrightarrow{V_c}$. Dos modificaciones emplean solo el verbo ancla o la palabra candidata como únicos elementos tomados en cuenta de la expresión fija. La última modificación toma el vector contexto como un vector representativo de toda la frase F, mediante el promedio del los vectores de las palabras de la frase F ($\frac{\overrightarrow{V_c}+\overrightarrow{E_f}}{|F|}$) sin excluir ninguna palabra.

Después de generar nuestras listas ordenadas con nuestro método propuesto y con las tres variantes para su comparación, hemos etiquetado manualmente las primeras 100 frases y las últimas 100 de cada ordenamiento, para verificar la cantidad de frases fijas al principio y final de cada lista ordenada.

Las frases están ordenadas por la probabilidad de contener una expresión fija utilizando alguno de los métodos evaluados, por tanto, la calidad del método se refleja con la cantidad de expresiones fijas en las primeras posiciones del listado (nosotros examinamos 100) y la ausencia de expresiones fijas en las últimas posiciones.

Para medir la calidad del orden de las listas hemos utilizado métricas de evaluación provenientes de las tareas de recuperación de información (Manning et al., 2008). Medimos la precisión a 100 (P@100) que refleja la cantidad de frases con expresiones fijas en las primeras 100 posiciones; también evaluamos la precisión negativa a 100 (NP@100) que indica la cantidad de frases sin expresiones fijas en las ultimas 100 posiciones, y finalmente, la doble precisión (DP@100) es el producto de las dos métricas anteriores. Los resultados son mostrados en la tabla 1, para todas las métricas empleadas el mejor resultado ideal es 1.00 y el peor resultado 0.00, pero este escenario no suele presentarse en un caso real, ya que, en ocasiones, no hay suficientes expresiones fijas en el

corpus que contengan el verbo ancla para poder recuperar 100 frases correctas en las primeras posiciones.

Opacidad semántica	P@100	NP@100	DP@100
Método propuesto: $O_s(\overrightarrow{E_f} \cdot \overrightarrow{V_c})$	**0.17**	**0.94**	**0.16**
Énfasis en verbo-contexto: $O_s(\overrightarrow{v_k} \cdot \overrightarrow{V_c})$	0.02	**0.94**	0.02
Énfasis en verbo-expresión: $O_s(\overrightarrow{E_f} \cdot \overrightarrow{v_k})$	0.05	0.92	0.05
Énfasis en Frase-expresión: $O_s\left(\overrightarrow{E_f} \cdot \frac{\overrightarrow{V_c} + \overrightarrow{E_f}}{\lvert F \rvert}\right)$	0.16	0.90	0.14

Tabla 1: Resultados de la evaluación de la recuperación de frases fijas en el corpus de periódicos mexicanos.

En la tabla 1 podemos observar que el método propuesto que compara la expresión fija conformada por el verbo y la palabra candidata es la mejor alternativa de uso de la aproximación computacional de los campos semánticos. También observamos que el mejor método solo recupero 17 frases, cuando al menos sabemos de la existencia de 26 frases fijas en el corpus, es decir, está perdiendo al menos 9 frases. Hay que resaltar que esta propuesta ha sido significativamente mejor que aquellas en las que se pone más énfasis en el verbo (renglón 2 y 3); esto seguramente se debe a la naturaleza del verbo ancla, que al ser verbo soporte tiene muy poca carga semántica que pueda ser comparada por la medida de opacidad. En la renglón 5 vemos el método que emplea la frase completa, en lugar de solamente el contexto, con un resultado más cercano al método propuesto; pero podemos ver que este método deja muchas más expresiones fijas en el fondo de la lista, lo cual no es deseable porque típicamente estas no podrían revisarse para incorporarlas al diccionario.

Algunos ejemplos interesantes de las expresiones fijas en las frases encontradas dentro de las 100 posiciones se muestran en la tabla 2 junto con la posición en la que se encontraron.

Expresión fija	Ranking
“te da gusto”	3
“se dieron a la huida”	23
“dio voltereta”	26
“dieron lo mejor de sí mismo”	27
“da para mucho más”	48
“No me doy por vencido”	50
“Lo estoy dando por muerto”	50
“nos dan atole con el dedo”	51

Tabla 1: Algunos ejemplos bien recuperados de las expresiones fijas encontradas en el corpus de periódicos mexicanos con la posición en la que fueron encostradas.

Muchas de estas frases como "dado por muerto" son expresiones que se considerarían en proceso de fijación, lo cual es interesante ya que el método podría predecir cuáles son las expresiones que en los próximos años pueden empezar a fijarse en el lenguaje.

Al definir que el método propuesto tenía buenas prestaciones en la recuperación de expresiones fijas, creamos otra herramienta basada en el mismo método que permite determinar el uso opaco o transparente de las palabras en una frase, la cual es presentada en la siguiente sección.

2.2 DETECCIÓN DE EXPRESIONES FIJAS

El objetivo de esta segunda herramienta es predecir cuándo una frase contiene una expresión fija. Es sumamente útil conocer si una frase contiene una expresión fija tanto para la traducción como para cualquier proceso automático o de análisis manual.

Para el cálculo de la probabilidad de que una frase contenga una expresión fija usamos el mismo método que ha resultado efectivo en la primera herramienta presentada, es decir, la medida de opacidad semántica.

La detección en una frase se realiza mediante el uso de un umbral de confianza en la medida de opacidad semántica. Todo valor de opacidad semántica superior al umbral, se considera que es suficientemente alta como para indicar que las palabras evaluadas son parte de una expresión fija.

2.2.1 Experimentos

En un corpus de periódicos españoles se extrajeron todas las frases que contuvieran las palabras de 4 expresiones fijas bien conocidas. Cabe hacer notar que no en todas las frases las palabras fueron

usadas como parte de una expresión fija y, por tanto, la tarea de la detección es determinar en cada frase si las palabras son empleadas en una frase fija o semifija.

La detección del uso de una expresión fija en una frase es vista como una clasificación binaria (contiene/no contiene expresión fija) por lo tanto usamos las métricas de evaluación de la clasificación; el recuerdo, la precisión y el f-mesure (Manning et al., 2008).

Los resultados en la clasificación de las frases por cada una de las 4 expresiones fijas elegidas son mostradas en la tabla 3.

En la tabla 3 se puede observar que las frases fijas "abrir boca" y "cerrar boca" han sido más fáciles de identificar y también han sido expresiones fijas más populares que las otras, ya que estas frases suelen tener un uso más opaco que las otras expresiones evaluadas; este fenómeno indicaría que es más fácil de identificar las frases que son mayoritariamente opacas.

Expresión fija	Precisión	Recuerdo	F-Measure	Empleo Opaco
"abrir boca"	0.774	0.823	0.768	**0.823**
"cerrar boca"	**0.818**	**0.818**	**0.818**	0.818
"abrir ojo"	0.466	0.453	0.446	0.547
"cerrar ojo"	0.556	0.558	0.555	0.523

Tabla 3. Evaluación (Precisión, Recuerdo y F-Measure) de la detección del uso de la expresión fija y en la última columna porcentaje del uso opaco de la frase fija.

En general todas las métricas son consistentes y tienen una alta correlación con la frecuencia del uso opaco del término; esto revela que el método funciona mejor cuando se tienen suficientes ejemplos opacos de las frases para poder ajustar correctamente el umbral. Y hemos obtenido una precisión y recuerdo superior a 0.8 lo cual indica buen nivel de desempeño general.

3 PREDICCIÓN DE INTENCIÓN COMUNICATIVA Y POLARIDAD

Dentro del trabajo realizado en la estancia, se investigó el uso de algunas expresiones fijas y semifijas del francés en la expresión informal y espontánea, así como su relación con la comunicación

objetiva y subjetiva. Se realizaron experimentos para emplear un conjunto de expresiones fijas y semifijas en la predicción de la intención comunicativa y la polaridad de las frases subjetivas.

El método propuesto emplea una *base* de expresiones del francés (aproximadamente 800) que cuentan con un etiquetado de la intención comunicativa y de la polaridad (Hajok y Meneses Lerin, 2014). El método propuesto calcula la probabilidad de una frase mediante la suma del valor de polaridad de todas sus palabras que se encuentren dentro de los lemas de las expresiones de la base, como se indica en la ecuación 4.

$$Polaridad(F) = \sum_{w_i \in F} \begin{cases} Polaridad_{base}(w_i) & w_i \in base \\ 0 & w_i \notin base \end{cases} \quad (4)$$

La predicción de la intención comunicativa se realiza mediante tres umbrales dentro del valor del cálculo de la polaridad, por debajo del umbral negativo la frase se considera subjetiva con polaridad negativa; entre el umbral negativo y el umbral positivo se considera una frase neutra y por tanto asumimos que es objetiva, y finalmente, por encima del umbral positivo se considera como una frase subjetiva positiva.

3.1 EXPERIMENTOS

Se realizaron experimentos sobre el corpus de *tuits* de entrenamiento del taller-seminario DEFT 2015[1] que está enfocado en la detección y análisis de sentimientos en la expresión de medios sociales digitales (*tuiter*) en francés. Este corpus elegido cuenta con múltiples etiquetados de las frases (*tuits*), que permiten evaluar la capacidad del método propuesto para la predicción de la intención comunicativa y de la polaridad.

Como marco de evaluación del método usamos una división 80-20 que divide el corpus del DEFT 2015 (*training-section*) en 80% para entrenamiento, en donde se ajustan los valores óptimos de los umbrales, y el 20% para evaluación. Las métricas de evolución son las empleadas en las tareas de clasificación, F-Measure y Precisión promedio de las tres clases objetivo/neutro, positivo y negativa subjetivo.

[1] https://deft.limsi.fr/2015/corpus.en.php?lang=en

Comparamos el método propuesto con el uso de 2 métodos del estado del arte: el primer método es la aplicación de la Bolsa de palabras que es una técnica básica de minería de texto que emplea una representación de los textos basada en la aparición de las palabras que contienen los *tuits*. Se emplea un algoritmo de aprendizaje para identificar los ejemplos de entrenamiento y cuáles son las palabras más representativas de cada clase. El segundo método es el uso de una base ontológica SentiWordNet[2] (Baccianella et al., 2010) que cuenta con un valor de subjetividad positiva, subjetividad negativa y objetividad de cada palabra. Desafortunadamente esta base solo cuenta con palabras en Inglés, porque se tuvieron que traducir automáticamente las palabras de los *tuits* franceses a términos en Inglés, usando traducción de palabra por palabra en *Google traslate*[3]. Esta traducción introduce ciertos errores, pero es necesaria para poder utilizar el conocimiento del contenido en SentiWordNet. Finalmente, con la suma de la subjetividad positiva, negativa y de la objetividad de las palabras de los *tuits*, se emplea un algoritmo de aprendizaje para calcular, a partir de los ejemplos de entrenamiento, el umbral y confianza de cada uno de estos tres valores para determinar si el *tuit* es principalmente objetivo o subjetivo, con inclinación positiva o negativa.

Método	Precisión	F-Measure
Bolsa de Palabras	69.5402	0.686
Senti-WordNet	47.7011	0.458
Propuesto con base francesa	53.5104	0.466

Tabla 4. Resultados de la evaluación del método propuesto que emplea una base de expresiones francesas y de dos métodos del estado del arte.

Los resultados del método propuesto y de los dos métodos del estado del arte implementados para la comparación se muestran en la tabla 4.

De la tabla 4 resulta evidente que la bolsa de palabras es el mejor método para la predicción de la polaridad y la intención comunicativa. También encontramos que dentro de los métodos basados en bases de conocimiento, el método propuesto sustentado en la base francesa sobrepasa significativamente al método que usa *Senti-WordNet,* a pesar de que la base francesa es cerca de 35 veces menor. Este hallazgo también indica que la traducción directa introduce muchos errores en el

[2] http://sentiwordnet.isti.cnr.it/
[3] https://translate.google.com/

cálculo, debido seguramente a la incapacidad de reconocer los significados especiales de las palabras como las frases fijas.

4 CONCLUSIONES

El trabajo desarrollado en las estancias de investigación permitieron desarrollar tres aportaciones en el campo del estudio de las expresiones fijas en español y francés y su futura coordinación. Se obtuvieron dos herramientas útiles en la creación de los diccionarios monolingües de expresiones fijas y semifijas y se analizó lo explícito de las expresiones etiquetadas en la predicción de subjetividad y polaridad con resultados que muestran la importancia de considerar las expresiones propias del lenguaje.

5 REFERENCIAS

(Aguila Escobar, 2006). Aguila Escobar, Gonzalo "Las nuevas tecnologías al servicio de la lexicografía: los diccionarios electrónicos", Actas del XXXV simposio internacional de la sociedad española de lingüística. León, Universidad de León, 2006 ISBN: 84-690-3382-2. [http://fhyc.unileon.es/SEL/actas/Aguila.pdf consultado 18-05-2015]

(Papadopoulou, 2010). Papadopulou, Eleni, "Diccionario monolingüe coordinado para enseñanza/aprendizaje del griego moderno por parte de hispanohablantes y para traducción automático griego - español", Tesis doctoral Universidad Autónoma de Barcelona, Departamento de Filología Francesa i Románica, 2010.

(Gross, 1996), Gross, Gaston. *Les expressions figées en français: noms composés et autres locutions.* .Francia: Ophrys, 1996

(Stock y Stock, 2013), Stock, Wolfgang G. y Stock, Mechtild, "Handbook of Information Science", Ed. Walter de Gruyter Saur, 2013.

(Meneses Lerin, 2013). Meneses Lerin, José Luis, "El diccionario electrónico monolingüe coordinado el verbo dar Francés-Español (España) - Español (México)", Tesis doctoral Université Paris 13 (Sorbonne Paris Cité), Lexiques, Dictionnaires, Informatique, 2013.

(Manning et al., 2008). Manning, Christopher D., Raghavan, Prabhakar y Schutze, Hinrich., "Introduction to Information Retrieval", Cambridge University Press, 2008.

(Hajok y Meneses Lerin, 2014). Hajok, A. y Meneses Lerin, L. "FIESTA : Calcular la subjetividad en los textos de prensa franceses Voz "francófona Simposio" después de 2000", Adam Mickiewicz University, Poznan, Polonia, 20-21 marzo 2014.

(Baccianella et al., 2010). Baccianella, Stefano, Esuli, Andrea y Sebastiani, Fabrizio., "SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Minin", Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), 2010.