Análisis de una aplicación multilingüe del agrupamiento de textos

Alisa Zhila y Alexander Gelbukh

Centro de Investigación en Computación,
Instituto Politécnico Nacional, México, D.F.
alisa.zhila@gmail.com, www.gelbukh.com

Resumen Muchas tareas del procesamiento de lenguaje natural, tales como la traducción automática, la desambiguación de los sentidos de las palabras, y la detección de textos traducidos, entre otras, requieren del análisis de la información contextual. En el caso de los enfoques del aprendizaje automático supervisado, este análisis se debe previamente llevar a cabo por expertos humanos, lo cual es prohibitivamente costoso. Los enfoques del aprendizaje automático no supervisado ofrecen métodos totalmente automáticos para la resolución de las mismas tareas. Sin embargo, estos métodos no son robustos; sus resultados dependen mucho de los parámetros elegidos y son difíciles de interpretar. El agrupamiento de textos es una técnica no supervisada para el análisis de similitudes entre los textos. En este trabajo exploramos la utilidad del agrupamiento de textos para la detección de traducción de una palabra. Específicamente, analizamos cómo los resultados del agrupamiento dependen de los parámetros del agrupamiento y evaluamos los resultados comparándolos con las traducciones hechas por traductores humanos.

Palabras clave: Agrupamiento de textos, traducción automática, desambiguación en la traducción, métodos no supervisados, selección de parámetros, contexto.

1. Introducción

En el procesamiento de lenguaje natural [16] la tarea de la traducción automática es uno de los problemas más importantes de este campo. La tarea de detección de traducción de una palabra se considera como un caso particular de la traducción automática. Esta tarea consiste en la selección de la mejor traducción, o varias traducciones equivalentes (sinónimas), para una palabra polisémica en un contexto dado, de un conjunto de variantes de traducción ofrecido por el diccionario bilingüe.

En la última década fueron explorados los enfoques para el mejoramiento de la traducción automática con la desambiguación de los sentidos de las palabras efectuada antes de la traducción [1, 14]. Estos enfoques se basan en los clasificadores supervisados, los cuales requieren un amplio entrenamiento con un corpus etiquetado manualmente. Requieren de muchos recursos computacionales pero proporcionan una mejora relativamente poca a un costo alto.

La tarea de la desambiguación de traducción, también llamada detección de la traducción, fue considerada en [4, 15] con recursos costosos, ya sea con un clasificador supervisado o con un corpus monolingüe etiquetado muy grande.

El agrupamiento de textos es un enfoque no supervisado para la detección de similitud entre los textos [6, 10]. Sin embargo, sus resultados dependen en gran medida de los parámetros utilizados para el agrupamiento. Este enfoque se aplicó a la tarea de discriminación de los sentidos de las palabras [9], la cual consiste en detección de la existencia de diferentes acepciones de la palabra sin que éstas estén previamente especificadas en algún diccionario.

La aplicación de los enfoques no supervisados a la detección de la traducción de una palabra no se ha explorado todavía. Nuestra hipótesis es que el agrupamiento no supervisado de textos aplicado a un corpus de textos paralelos alineados por palabras puede ser útil para obtener las características del contexto que permitan la selección correcta de una variante de traducción para una palabra en un contexto dado de manera no supervisada.

En este trabajo exploramos la utilidad del agrupamiento de textos para la detección de la traducción mediante la comparación de los resultados del agrupamiento obtenidos con varias combinaciones de los parámetros. Evaluamos los resultados obtenidos comparándolos con las traducciones humanas obtenidas del corpus de textos paralelos alineados.

El resto del artículo está estructurado de la siguiente manera. En la sección 2 se explican los conceptos del agrupamiento de texto. En la sección 3 se dan a conocer los parámetros de los algoritmos que hemos explorado en este trabajo. En la sección 4 se presentan los resultados obtenidos. Finalmente, la sección 5 presenta una discusión de las lecciones aprendidas y las conclusiones.

2. El agrupamiento de los textos

En la última década el tema de la discriminación no supervisada de los sentidos de las palabras (WSD por sus siglas en inglés: *word sense discrimination*), o sea, la distinción entre los usos diferentes de una palabra en contextos diferentes, se ha investigado activamente. La solución más conocida a este problema es el agrupamiento de los contextos que contienen la palabra en cuestión [6, 10]. Una revisión amplia del agrupamiento como la clasificación no supervisada de los elementos del conjunto de datos en grupos se presenta en [2]. Los algoritmos de agrupamiento adecuados para el agrupamiento de documentos se describen y analizan en [11]. Estos algoritmos son implementados en la herramienta de agrupamiento que se llama SenseClusters [8]. Nosotros adoptamos este programa como la herramienta central para nuestros experimentos.

Sin embargo, los resultados del agrupamiento de textos dependen en gran medida de los parámetros con los cuales se realiza el agrupamiento. En esta sección presentamos una breve descripción de los parámetros y las técnicas del agrupamiento.

2.1. Los tipos de las características de los textos

Para llevar a cabo un agrupamiento de elementos, antes de todo es necesario elegir características para representar cada elemento. En el campo del agrupamiento automático de los textos, las características son usualmente secuencias de palabras,

llamadas *n*-gramas. Los *unigramas* son las características que se forman de una sola palabra. Pares de palabras consecutivas se denominan *bigramas*. Una definición extendida en [3] propone que bigramas son pares de palabras que se encuentran dentro de cierta distancia (*ventana*) una de otra, preservando el mismo orden en el cual aparecen en el texto. Los pares de palabras dentro de una ventana sin importancia del orden se llaman *coocurrencias*. Para los textos que contienen una palabra específica cuyas propiedades están bajo el estudio, como en el caso de la discriminación de los sentidos de la palabra dada, se consideran *coocurrencias con la palabra objetivo* (es decir, la palabra dada), las cuales son las coocurrencias que contienen la palabra en cuestión.

Las características que se encuentran un número de veces menor de un *umbral r* (el parámetro de corte de la frecuencia) no pueden servir como una base sólida para el agrupamiento de textos y, por lo tanto, se excluyen de la lista de las características.

2.2. El orden de la representación de los textos

Se consideran representaciones de primer y de segundo orden. La representación de primer orden representa un texto como un vector solamente de las características que se encuentran directamente en el texto. La representación de segundo orden considera también las características que se encuentran junto con las características del texto dado, en otros textos existentes.

Por ejemplo, si el texto A es "ratón de computadora", sus unigramas de primer orden son "ratón" y "computadora". La preposición "de" se elimina ya que aparece en una lista de palabras basura (*stopwords* en inglés), las cuales son las palabras que no agregan información específica sobre el texto. En el caso de la presentación de segundo orden, considerando también el texto B "ratón inalámbrico", las características del texto A resultan ser "ratón", "computadora" e "inalámbrico". De ese modo se aumenta la cantidad de las características en casos de los textos cortos.

2.3. Medidas de similitud

Para evaluar la similitud entre los textos, es necesario introducir una medida de similitud correspondiente a la representación seleccionada de las características. Usualmente los elementos son representados como vectores de características, con lo cual tales se pueden utilizar tales medidas de similitud entre los vectores como distancia o coseno. Un conjunto de textos se puede representar en un espacio vectorial, donde un vector corresponde a un documento, o bien se puede construir una matriz de similitudes entre pares de textos.

2.4. Funciones de criterio de agrupamiento

La tarea del agrupamiento de elementos es la optimización de una función llamada la función de criterio de agrupamiento (*clustering criterion function* en inglés), la cual es una función de medida de similitud. Una revisión y comparación de las funciones de criterio de agrupamiento se presenta en [6], donde los autores comparan varias funciones de criterio de agrupamiento. Las funciones de criterio internas denominadas

I1, I2 e I3 [6] se basan en la similitud dentro de un grupo de elementos, mientras que las funciones de criterio externas denominadas E1 y E2 toman en cuenta las distancias entre los grupos. Las funciones denominadas H1 y H2 son híbridas: combinan las propiedades de las funciones de criterio internas y externas. Todas estas funciones consideran un documento o texto como un vector de características. En [6] se muestra que las funciones I2 y H2 dan los mejores resultados con la mayoría de los algoritmos de agrupamiento.

2.5. Técnicas de agrupamiento

Existen diversas técnicas, o algoritmos, de agrupamiento. Éstas se dividen en dos grandes categorías: jerárquicas y de partición. La descripción detallada de las técnicas de agrupamiento y la comparación de sus aplicaciones para el agrupamiento de textos se presenta en [11]. Esa investigación mostró que las mejores técnicas de agrupamiento de textos son las siguientes: *k*-medias (*k-means* en inglés), la cual es una técnica de partición con *k* centros de agrupamiento, y el agrupamiento aglomerativo con *UPGMA* (por sus siglas en inglés, *Unweighted Pair Group Method with Arithmetic mean*), el cual es un algoritmo jerárquico.

2.6. Criterios de paro del agrupamiento

Las técnicas de agrupamiento existentes requieren que el número de los clústeres resultantes sea conocido de antemano, lo cual no se cumple en muchos casos prácticos, especialmente para el agrupamiento de textos y la discriminación de los sentidos de las palabras. Para los casos cuando el número deseado de grupos no está conocido de antemano, en [3] se introducen los criterios de paro automático del proceso de agrupamiento. El criterio *gap* se basa en la llamada estadística GAP aplicada a la dispersión de los elementos dentro de cada grupo. Los criterios PK1, PK2 y PK3 (PK refiere a "*P*redecir el número *K* de grupos") miden el valor del cambio de la función de similitud de cada grupo con el número *k* de los grupos.

En este trabajo se utilizan los criterios de paro de agrupamiento para detectar automáticamente el número de sentidos del término en cuestión, ya que la idea de la discriminación de sentidos de palabra sostiene que los usos similares de una palabra se encuentran en los contextos similares. A continuación, comparamos el número resultante de los grupos con el número de acepciones proporcionadas en los diccionarios, para evaluar el desempeño de los criterios de paro automático del agrupamiento.

2.7. Evaluación del agrupamiento

La evaluación del desempeño de los algoritmos de agrupamiento depende considerablemente de la medida que se utiliza para evaluarlo [11]. Hay dos enfoques básicos para la evaluación del agrupamiento: interno y externo. Las medidas internas no utilizan ningún conocimiento externo sobre las posibles soluciones (clases de elementos). Las medidas externas comparan los resultados de agrupamiento contra las clases de los elementos conocidas con algún procedimiento externo.

Nosotros evaluamos los resultados del agrupamiento comparándolos con el conjunto de las variantes de traducción obtenidos con un corpus de textos paralelos, véase la sección 3. Elegimos las medidas externas de la entropía y la pureza, cuyas definiciones y fórmulas las adoptamos del trabajo [11].

La entropía es la medida de la incertidumbre de la distribución de las clases por los grupos. Para un grupo S_r con n_r elementos, la entropía del grupo es

$$E(S_r) = -\frac{1}{\log q} \sum_{i=1}^{q} \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r},$$

donde q es el número de las clases y n_r^i es el número de los elementos de la clase i que se encuentran en el grupo r. La entropía resultante es la sumatoria ponderada de las entropías de todos los clústeres:

$$E(S_r) = \sum_{r=1}^{k} \frac{n_r}{n} E(S_r), \tag{1}$$

donde k es el número de los grupos y n es el número de todos los elementos en el conjunto de datos.

La pureza de un cluster se calcula como:

$$P(S_r) = \frac{1}{n_r} \max_i (n_r^i).$$

Es la fracción más grande de un grupo formada por los elementos de una clase, es decir, es la fracción de la clase más grande en un grupo.

La pureza total de un conjunto de grupos es la sumatoria ponderada de todas las purezas individuales de los clústeres en el conjunto:

$$P = \sum_{r=1}^{k} \frac{n_r}{n} P(S_r).$$

La pureza evalúa la calidad con la cual cada clase corresponde a un grupo. En resumen, la entropía más baja y la pureza más alta indican mejor calidad del agrupamiento.

3. Metodología experimental

En esta sección describimos los experimentos llevados a cabo en este trabajo. Nuestro objetivo es la exploración de la utilidad del agrupamiento no supervisado de textos para la detección de la traducción. Para este experimento hemos utilizado la herramienta existente SenseCluster [8] para realizar el agrupamiento de los textos de nuestra base de textos.

Nuestros experimentos se basan en los resultados obtenidos en [7, 9, 11, 13]. Exploramos si estos parámetros de agrupamiento son apropiados para la detección de la traducción. Nuestra hipótesis principal es que las traducciones son correlacionadas con las acepciones de la palabra que se traduce.

La discriminación no supervisada de los sentidos de las palabras sostiene que los usos similares con las mismas acepciones se agrupan. Por lo tanto, los contextos de un grupo corresponderían a una sola traducción o a las traducciones sinónimas. Utilizamos los criterios de paro del agrupamiento para detectar automáticamente el número de acepciones de la palabra en cuestión.

A continuación, usamos el número de acepciones proporcionadas por los diccionarios como el umbral para el número de los grupos.

3.1. El corpus de textos

Extrajimos nuestro corpus de textos para el experimento del corpus paralelo inglés-español Europarl disponible con el "corpus abierto OPUS" [12].

Para nuestro propósito de explorar la posibilidad del uso del agrupamiento de textos para la detección de la traducción, una palabra ambigua tenía que satisfacer los siguientes criterios:

– ser encontrada en el corpus paralelo al menos mil veces, lo cual es suficiente para llevar a cabo el agrupamiento de textos no supervisado;
– tener más de una variante de traducción en la parte paralela del corpus.

Hemos encontrado las siguientes palabras en la parte inglesa del corpus que satisfacían estos dos criterios: *facility*, *post*, *language*. Debido a las limitaciones del espacio, en este artículo sólo presentamos los resultados para la palabra *facility*.

Para formar un contexto extrajimos 7 oraciones consecutivas del corpus de manera tal que una de ellas contenga la palabra elegida. Elegimos el tamaño de 7 oraciones basándonos en los tamaños promedios de los contextos utilizados en [7, 9, 11, 13]. En este paso, extrajimos 1771 contextos.

Luego eliminamos los contextos mal alineados de nuestro conjunto experimental de textos. Los casos en los cuales la palabra *facility* no tenía un equivalente de traducción los marcamos como `NOTAG` (palabra no etiquetada). Agrupamos las variantes de las traducciones de baja frecuencia (de 1 a 6 ocurrencias en el corpus) con sus sinónimos, prestando atención a sus usos en contexto. Finalmente, obtuvimos un conjunto de 1429 contextos y 21 clases de traducción, incluyendo la clase `NOTAG`[1]. Estas clases de traducción sirven para la evaluación externa de los clústeres resultantes.

Los diccionarios monolingües que hemos consultado (Merriam-Webster en línea, *Oxford Concise Thesaurus*, WordNet y *Larousse American Pocket*) presentan entre 4 y 5 acepciones para la palabra *facility*. Nosotros adoptamos el menor de estos números (4) como el umbral para el número mínimo de los grupos. Por eso,

[1] El conjunto de los textos está disponible en http://www.gelbukh.com/resources/word-translation-alignments/

eliminamos cualquier combinación de valores de parámetros que genera menos de 4 grupos.

3.2. Los parámetros del agrupamiento

En este trabajo llevamos a cabo el agrupamiento de textos con la herramienta SenseClusters [8]. Es un sistema para el agrupamiento de textos completo y libremente disponible, el cual proporciona las posibilidades para la selección de las características de los textos, varios esquemas de representación de éstos, los algoritmos de agrupamiento diferentes y la evaluación de los grupos obtenidos.

Los parámetros con valores fijos. Los valores de los siguientes parámetros no se variaron en nuestros experimentos:

- el orden de la representación: usamos la representación de segundo orden (*O2*);
- los textos fueron representados como vectores de características en el espacio vectorial;
- el ancho de la ventana entre las palabras de características fue 5;
- el parámetro *r* de la frecuencia de corte fue 3.

Elegimos la representación de los contextos de segundo orden ya que se ha demostrado que esta representación es mejor para los contextos cortos [5]. El espacio vectorial es preferible a la matriz de similitud según el trabajo [9]. Para el valor del ancho de la ventana tomamos como referencia el trabajo [7]. Hemos establecido el valor del parámetro *r* de la frecuencia de corte a 3 heurísticamente, basándonos también en el trabajo [7].

Los parámetros con los cuales experimentamos. Ya que el número total de las posibles combinaciones de los valores de los parámetros es muy alto, analizamos sólo los valores de los parámetros que se han demostrado ser mejores para el agrupamiento de textos [11, 13]. Los parámetros con los valores que variamos eran los siguientes:

- las características para la representación de contexto: unigramas, bigramas, coocurrencias y coocurrencias con la palabra objetiva;
- las técnicas del agrupamiento: k-medias, bisección repetida, bisección repetida refinada y el agrupamiento aglomerativo;
- las funciones de criterio de agrupamiento: I2, H2 y UPGMA;
- los criterios de paro del agrupamiento: *gap*, PK1, PK2, PK3.

El número total de los experimentos era 112.

4. Resultados experimentales

El número de los grupos que obtuvimos con diversas combinaciones de los parámetros del agrupamiento variaba de 1 a 6.

Los criterios de paro del agrupamiento. La tabla 1 muestra las frecuencias de cada número de los grupos para los criterios de paro del agrupamiento.

Tabla 1. Número de grupos obtenido con cada criterio de paro.

	Número de grupos					
	debajo del umbral			arriba del umbral		
Criterio	1	2	3	4	5	6
gap	24	0	4	0	0	0
PK1	11	10	3	1	3	0
PK2	0	8	10	3	4	3
PK3	0	12	9	6	1	0

Como se observa en la tabla 1, los criterios de paro del agrupamiento *gap* y PK1 resultan en el menor número de grupos. El criterio *gap* no ha producido el número de grupos más del umbral de 4. El criterio PK1 dio resultados aceptables sólo en 4 casos, lo cual representa el 3.5% de los casos.

La proporción del número de experimentos por cada cantidad de los grupos obtenidos se muestra en la tabla 2.

Tabla 2. Proporción del número de experimentos por cantidad de los grupos obtenidos.

número de grupos	1	2	3	4	5	6
proporción de experimentos, %	31.2	26.8	23.2	9.0	7.1	2.7

El los 50% de la cantidad total de los experimentos fueron obtenidos de 2 a 3 grupos, y sólo el 18,8% (21 del total de los 112 experimentos) pasaron sobre el umbral de 4 grupos.

La suposición de que la palabra *facility* puede tener sólo 2 a 3 acepciones reales o bien distinguidos no parece sólida. Analizando las definiciones de *facility* en los diccionarios se puede observar que sus acepciones difícilmente se puedan agrupar en la cantidad de acepciones independientes y semánticamente no relacionadas menor de 4 acepciones.

Por lo tanto, interpretamos las cantidades bajas de los grupos que se produjeron con los criterios de paro del agrupamiento *gap* y PK1 como una cualidad inherente a estos criterios. Los criterios PK2 y PK3 dieron resultados aceptables en el 36% y el 25% de los casos, respectivamente.

Características de contextos, la entropía y la pureza. Para cada experimento con el número de clústeres resultantes mayor que el umbral de 4, los valores de los parámetros de la entropía y la pureza correspondientes se presentan en la tabla 3. Las columnas en esta tabla son las siguientes: *técnica* refiere a la técnica de agrupamiento, *criterio* refiere a la función del criterio de agrupamiento, *paro* refiere al criterio de paro del agrupamiento, *grupos* es el número resultante de los grupos para la combinación dada de los parámetros y E y P son la entropía y la pureza, respectivamente. Las abreviaturas son como sigue: *agglo* para la técnica de agrupamiento aglomerativo, *direct* para k-medias, *br* para bisección repetitiva, *brr* para bisección repetitiva refinada. Estas abreviaturas se utilizan en el resto del artículo. Otros términos se han explicado en las secciones 2 y 3.2.

Ninguno de los experimentos con las bigramas o las coocurrencias con la palabra objetiva resultaron en el número de grupos mayor o igual a 4. Por lo tanto, los

resultados de todos los experimentos con estas características se descartaron y no se muestran en la tabla 3.

Tabla 3. Resultados obtenidos con diferentes técnicas de agrupamiento.

técnica	criterio	paro	grupos	E	P
Coocurrencias					
agglo	upgma	pk2	6	80.6	25.5
direct	h2	pk1	4	80.4	25.6
direct	h2	pk3	4	80.4	25.6
direct	i2	pk2	5	80.2	25.5
direct	i2	pk3	4	80.4	25.6
br	h2	pk1	5	80.7	25.0
br	h2	pk2	4	81.0	25.0
br	h2	pk3	4	81.0	25.0
br	i2	pk3	5	80.7	25.0
brr	h2	pk3	4	80.4	25.6
brr	i2	pk2	5	80.2	25.5

técnica	criterio	paro	grupos	E	P
Unigramas					
agglo	upgma	pk2	6	84.1	24.2
direct	i2	pk2	6	**74.8**	26.9
br	h2	pk1	5	75.2	**28.3**
br	h2	pk2	4	76.2	27.6
br	h2	pk3	4	76.2	27.6
br	i2	pk2	5	75.6	27.8
brr	h2	pk1	5	75.2	**28.3**
brr	h2	pk2	4	76.2	27.6
brr	h2	pk3	4	76.2	27.6
brr	i2	pk2	5	75.3	**28.3**

En la tabla 3 los mejores valores de la entropía y la pureza se dan en negritas. Todos ellos fueron obtenidos para las combinaciones de los parámetros con las unigramas. Generalmente, la entropía y la pureza para las unigramas son aproximadamente 5% y 12% mejores, respectivamente, que las de las coocurrencias. Sin embargo, la comparación de los valores de la entropía y la pureza con los obtenidos en [11, 13] se obstaculiza porque la entropía y la pureza dependen del número de las clases externas.

Podemos deducir que la entropía y la pureza como se describen en [11] podrían no ser apropiados para nuestra tarea. En [11], estas medidas se utilizan para evaluar los resultados de la discriminación de los sentidos de las palabras cuando se supone que cada grupo corresponde a una acepción de la palabra y se supone que el número de los grupos sería igual al número de las acepciones, que son las clases externas usadas para la evaluación. En nuestro caso, es perfectamente aceptable que más de una clase se agrupa en un grupo o que los elementos de una clase se distribuyen entre varios grupos, lo cual es interpretable desde el punto de vista de la traducción.

Número de grupos. Para investigar la influencia del número de los grupos en la entropía y la pureza, llevamos a cabo un experimento con un número de grupos fijo establecido de forma manual a 21, que es el número de las clases de traducción. En este experimento se utilizó una combinación de parámetros de agrupamiento que produjo la pureza más alta. Los resultados se muestran en la tabla 4.

Como se observa, aumentando el número de los grupos más de cuatro veces (de 5 a 21) se mejoran los valores de la entropía y la pureza sólo al 10.6% y 15.5%, respectivamente.

Tabla 4. Los resultados obtenidos con 21 grupos.

técnica	criterio	paro	grupos	E	P
Unigramas con el número de clústeres fijo					
rb	h2	n/a	21	67.2	32.7

Ejemplo de los resultados del agrupamiento. Para ilustrar los resultados del agrupamiento, hemos elegido dos experimentos que dieron los mejores resultados: uno para las coocurrencias y el otro para los unigramas. La tabla 5 muestra la distribución de las clases de traducción por los grupos resultantes para estas dos combinaciones de los parámetros.

En la sección 3 hemos explicado la suposición de que el agrupamiento no supervisado de textos sería apropiado para la detección de la traducción de la palabra si un grupo correspondiera a una o más clases enteras de traducción, lo que es el caso de sinonimia entre las traducciones, o bien si una clase de traducción fuera distribuida entre algunos grupos, lo que es el caso de la homonimia que se presenta en ambos idiomas. En estos casos, la tabla de la distribución de las clases por los grupos se vería más "diagonal", o bien tendría más celdas con ceros que las con no ceros. Sin embargo, en la tabla 5 observamos que casi todas las celdas tienen valores distintos de cero. Esto significa que un contexto correspondiente a cualquier variante de traducción se puede encontrar en cualquier grupo, lo que en este caso invalida nuestra hipótesis inicial acerca de la idoneidad del agrupamiento de textos para la detección de traducción.

Tabla 5. Ejemplo de la distribución de las clases por los grupos, en dos experimentos.

Experimento	Tamaño del grupo	*dispositivo*	*equipos*	*facilidad*	*instalación*	NOTAG	*servicio*	*sistema*	*posibilidad*	*mecanismo*	*institución*	*capacidad*	*centro*	*ayuda*	*medio*	*crédito*	*planta*	*fondo*	*medida*	*infraestruct*	*plan*	*central*
	Cooccurrencias, agrupamiento aglomerativo, UPGMA, PK2, E = 0.806, P= 0.255																					
0	989	42	22	121	196	94	96	19	95	82	16	25	18	19	55	18	16	20	8	14	4	9
1	3	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
2	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	350	3	6	24	142	40	40	3	14	6	10	4	9	0	20	1	6	1	5	2	0	14
4	11	0	0	5	1	1	0	0	2	0	0	0	1	0	0	0	0	0	0	1	0	0
5	75	18	1	1	4	3	5	19	4	8	2	0	1	2	1	1	0	5	0	0	0	0
	Unigramas, bisecciones repetitivas, H2, PK1, E = 0.758, P = 0.276																					
0	213	27	1	24	8	17	16	6	24	26	1	9	3	7	12	9	1	15	1	3	2	1
1	156	1	0	1	112	10	3	0	1	1	0	0	1	0	5	0	8	0	2	0	0	11
2	282	0	13	42	83	26	35	3	22	4	6	4	8	3	13	1	3	0	2	9	0	5
3	307	32	1	29	24	25	11	25	32	57	4	9	6	7	18	7	3	10	4	0	2	1
4	471	3	14	56	116	61	77	7	36	8	17	7	11	4	28	4	7	1	4	5	0	5

5. Conclusiones y trabajo futuro

En este trabajo hemos realizado una comparación de las combinaciones de los parámetros del agrupamiento de textos y exploramos su utilidad para la detección no supervisada de la traducción correcta de una palabra. Específicamente, presentamos los resultados para la palabra inglesa *facility* y sus traducciones al español.

Sólo el 18,8% de los experimentos resultaron en el número de grupos que sobrepasó el umbral de 4, que es el número mínimo posible de las acepciones de la palabra *facility*. Entre 2 y 3 grupos se detectaron en el 50% de los casos. En el trabajo actual estos resultados no se interpretan desde el punto de vista semántico y simplemente los excluimos de consideración. Sin embargo, el análisis formal de la similitud semántica de los sentidos a través de una ontología o de una jerarquía semántica puede dar una nueva perspectiva a estos resultados.

Hemos detectado que el uso del criterio de paro del agrupamiento *gap* en los experimentos resulta en el número de grupos muy escaso que no se puede interpretar desde el punto de vista semántico. Logramos obtener los números de los grupos que corresponden a la suposición semántica del número de acepciones de la palabra utilizando los criterios PK2 y PK3. Además, no eliminamos de la consideración el criterio PK1, ya que su uso proporciona el 19% de todos los resultados aceptables.

No hemos detectado resultados aceptables para las coocurrencias con la palabra objetiva y los bigramas. Eso podría explicarse por el valor del parámetro del ancho de la ventana inadecuado y por la dispersión de las características.

La evaluación de los resultados a través de la entropía y la pureza dio los números que no son fáciles de interpretar en el marco de la detección de traducción de una palabra, ya que el número de las clases es mucho mayor que el número de los grupos. Por lo tanto, planeamos trabajar en el desarrollo de una medida diferente del desempeño del agrupamiento la cual sería más adecuada para nuestros objetivos.

Agradecimientos. El trabajo presentado en este artículo fue apoyado parcialmente por el Gobierno de México a través, los proyectos CONACYT 50206-H y 122030 (CONACYT-DST India 2011–2014), SIP-IPN 20121823, así como el SNI y el programa PIFI-IPN.

Referencias

1. Carpuat, M., Wu, D.: Improving statistical machine translation using word sense disambiguation. In: Proc. of EMNLP-CoNLL 2007, pp. 61–72 (2007)
2. Jain, A.K., Murty, M.N., Patrick, J. Flynn, P.J.: Data Clustering: A Review. ACM Computing Syrveys, vol. 21, pp. 264-323 (1999)
3. Kulkarni, A., Pedersen, T.: Unsupervised Context Discrimination and Automatic Cluster Stopping. MS Thesis, University of Minnesota, UMSI 2006/90 (2006)
4. Marsi, E., Lynum, A., Bungum, L., Gambäck, B.: Word Translation Disambiguation without Parallel Texts. In: Proc. International Workshop on Using Linguistic Information for Hybrid Machine Translation, Barcelona, Spain (2011)
5. Pedersen T.: Computational Approaches to Measuring the Similarity of Short Contexts: A Review of Applications and Methods. University of Minnesota, UMSI 2010/118 (2008)

6. Pedersen, T., Bruce, R.: Distinguishing word senses in untagged text. In: Proc. of the 2nd Conference on Empirical Methods in Natural Language Processing, Providence, RI, pp. 197–207 (1997)
7. Purandare, A.: Unsupervised Word Sense Discrimination By Clustering Similar Contexts. MS Thesis. University of Minnesota. (2004)
8. Purandare, A., Pedersen, T.: SenseClusters – Finding Clusters that Represent Word Senses. In: Proc. of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-04), pp. 26-29 (2004)
9. Purandare, A., Pedersen, T.: Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces. In: HLT-NAACL 2004 Workshop, CoNLL-2004, pp. 41-48 (2004)
10. Schütze, H.: Automatic Word Sense Discrimination. Journal of Computational Linguistics. vol. 24(1), pp. 97-123 (1998)
11. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. University of Minnesota, Technical Report 00-034 (2000)
12. Tiedemann, J.: News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. Recent Advances in Natural Language Processing, vol. V, pp. 237-248 (2009)
13. Zhao, Y., Karypis, G.: Criterion Functions for Document Clustering: Experiments and Analysis. University of Minnesota, Technical Report 01-040 (2001)
14. Vickrey, D., Biewald, L., Teyssier, M., Koller, D.: Word-sense disambiguation for machine translation. In: Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing 2005, pp. 771–778 (2005)
15. Holmqvist, M.: Memory-based learning of word translation. In: Proc. of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007, Tartu, Estonia, pp. 231-234 (2007)
16. Ledeneva, Y., Sidorov, G.: Recent Advances in Computational Linguistics. Informatica. International Journal of Computing and Informatics, 34, 3–18 (2010)