# Redes de palabras alineadas como recurso en la extracción de equivalencias léxicas de traducción y su aplicación en la alineación

Eduardo Cendejas, Grettel Barceló, Gigori Sidorov, Alexander Gelbukh, and Liliana Chanona-Hernandez

***Resumen*—La equivalencia léxica de traducción se define mediante correspondencias establecidas entre dos lenguas, comúnmente denominadas lengua de origen y lengua meta. Este artículo propone un método de extracción de dichas equivalencias en palabras no funcionales. El algoritmo se basa en dos recursos principales: 1) MultiWordNet como léxico especializado para cada uno de los idiomas involucrados y 2) textos paralelos como información adicional para proporcionar diversas lexicalizaciones de las palabras a corresponder. Utiliza como fundamento principal el hecho de que las redes de palabras que conforman MultiWordNet están alineadas. Además, se presenta la reutilización del repositorio de pares léxicos obtenidos, señalando la forma en que esta información es susceptible de ser usada en un sistema de alineación a nivel de palabras. Para realizar los experimentos se emplearon textos paralelos bilingües sin notación morfosintáctica alguna, alineados a nivel de oración en los pares de idiomas español / inglés y español / italiano.**

***Palabras clave*—Equivalencias léxicas, alineación, redes de palabras, textos paralelos.**

## Aligned Word Networks as a Resource for Extraction of Lexical Translation Equivalents, and their Application to the Text Alignment Task

***Abstract*—The notion of lexical translation equivalent is defined via correspondence established between two languages conventionally called source and target languages. We propose a method for extraction such equivalents for non-functional words. Our algorithm is based in two main resources: (1) MultiWordNet as a specialized lexicon for each one of the two languages in question and (2) a parallel text corpus as a source of additional information that provides various lexicalizations of the words that are being aligned. Our method is based on the fact that the word networks that form MultiWordNet are aligned. In addition, we discuss an application of the obtained list of word pairs in a word-level text alignment system. In our experiments we used bilingual sentence-level aligned parallel texts, without any morphosyntactic annotation, for word pairs Spanish / English and Spanish / Italian.**

***Index Terms*—Translation equivalents, alignment, word networks, parallel texts.**

Manuscript received December 9, 2011. Manuscript accepted for publication March 8, 2012.

E. Cendejas, G. Barceló, G. Sidorov and A. Gelbukh are with the Center for Computing Research, National Polytechnic Institute, Mexico City, Mexico. (web: cic.ipn.mx/sidorov, www.gelbukh.com)

## I. Introducción

En la década de los ochentas se introdujo la idea de almacenar electrónicamente traducciones pasadas en un formato bilingüe. El concepto consistía en la construcción de una concordancia bilingüe teniendo un operador de datos que manualmente introdujera el texto y su traducción [1], creando así una herramienta de referencia valiosa para los traductores. Como consecuencia, se originó un gran interés en la construcción automática de tales bases de datos a mayor escala.

En años recientes, se ha suscitado un progreso considerable en el campo de la equivalencia y alineación paralela de textos. El creciente interés en éstos viene dado básicamente de la popularización de Internet que ha hecho de la red una enorme colección documental bilingüe. La expresión «texto paralelo» por sí misma es ahora bien establecida dentro de la comunidad de la lingüística computacional.

Un texto paralelo es la unión de dos o más textos que poseen el mismo contenido semántico, pero expresados en lenguajes diferentes [2]. El término paralelo no implica que los textos tengan una correspondencia exacta entre palabras, oraciones y/o párrafos; es decir, dos textos pueden estar completamente desalineados sin dejar de ser textos paralelos [3].

### *I-A. Equivalencia vs. alineación*

Uno de los problemas clásicos y tradicionales de la teoría de la traducción ha sido, y lo sigue siendo, el de la equivalencia. En esta tarea, un par de traducción es considerado correcto si existe al menos un contexto en el cual éste ha sido acertado. Usualmente la extracción sólo está interesada en las categorías principales (sustantivos, adjetivos y verbos).

Sin embargo, la alineación a nivel de palabras es diferente y más difícil que el problema de extracción de equivalencias [4]. Consiste en indicar qué palabra del lenguaje origen (LO), se corresponde con una palabra en el lenguaje meta (LM), hasta encontrar todas las correspondencias entre las palabras de los textos que conforman el corpus bilingüe [5]. Por tanto, requiere que cada palabra (sin importar el POS –*Part of speech*) o signo de puntuación en ambas partes del bitexto sean asignados a una traducción o a un nulo en su contraparte.

## II. Trabajos previos

Tanto la extracción de equivalencias léxicas, como la alineación de textos, han demostrado ser fuentes inestimables de datos de traducción para los bancos de terminología y los diccionarios bilingües. Actualmente, ambas tareas están proporcionando la base para el desarrollo de una nueva generación de herramientas de asistencia para los traductores humanos, que permitan mejorar la calidad y productividad de su trabajo.

Los métodos propuestos para encontrar equivalentes y alineaciones en textos paralelos, se han clasificado generalmente en dos tipos de aproximaciones: los estadísticos y los lingüísticos.

Los métodos *estadísticos* clásicos utilizan información no-léxica, como la correlación esperada de longitud y posición de las unidades de texto (párrafos u oraciones), la frecuencia de co-ocurrencia, la proporción del tamaño de las oraciones en los diferentes idiomas, etc. [6]. Intentan establecer la correspondencia entre las unidades del tamaño esperado [7], [8], [9], el cual puede medirse en el número de palabras o caracteres [10].

Por otro lado, los métodos *lingüísticos* se apoyan en recursos léxicos existentes, como diccionarios bilingües de gran escala y glosarios [6], para establecer la correspondencia entre las unidades estructurales. Por la disponibilidad cada vez mayor de recursos bilingües, se invierte más esfuerzo en la investigación de la efectividad de los acercamientos basados en léxicos [11], [12], [13].

La reestructuración que se realiza durante la alineación, independientemente del método que se requiera, puede ser realizada entre párrafos, sentencias o palabras, del contenido expresado en lenguaje original y su traducción [14].

Por otra parte, a pesar de que la extracción de equivalencia puede realizarse en cualquiera de los niveles de emparejamiento, del mismo modo que la alineación, ésta ha estado enfocada a la identificación de pares de palabras o secuencias de palabras (con patrones establecidos [15] o colocaciones [16]).

La correspondencia de los textos paralelos, en una primera fase a nivel de párrafo, es la más simple, pues casi siempre están en una relación 1:1, lo cual resulta fortuito ya que la correspondencia a nivel de oraciones mejora mucho si se realiza primero la relación a nivel de párrafos [3].

El nivel de resolución de oraciones presentó un desafío mayor, descubriendo que en él, las correspondencias uno a muchos y muchos a uno, no son raras. Se han publicado muchos algoritmos para alinear oraciones en textos paralelos: algunos se basan en la observación de la correlación entre la longitud de un texto y la de su traducción [7], [8], otros emplean técnicas estadísticas basadas en cognados [17] o que maximizan el número de las correspondencias sistemáticas entre las palabras [11], [18]. Aunque los resultados obtenidos por algunos de los métodos antedichos han sido absolutamente exactos cuando están probados en recopilaciones relativamente limpias y extensas, siguen siendo alineaciones "parciales", pues ocultan el grado más fino de resolución debajo del nivel de oración: el nivel de palabra.

La correspondencia a nivel de palabras tiene mayor dificultad que el de oración, puesto que la relación 1:1 llega a ser cada vez más rara. Se han propuesto varios métodos para encontrar equivalencias y alineaciones entre palabras en textos paralelos. Algunas técnicas dependen de un conjunto de parámetros que son aprendidos mediante un proceso de entrenamiento de datos [19], [20], [21]. Otros se basan en técnicas estadísticas, clasificadas principalmente en dos categorías [22]: métodos de prueba de hipótesis y métodos de estimación. Los primeros extraen los candidatos de equivalencia de las unidades de traducción y se someten a una prueba estadística, que miden la co-ocurrencia y/o similaridad de las cadenas [16], [23], [24]. Los enfoques de estimación hacen uso de modelos de alineación probabilística, calculados de corpus paralelos [25], [26]. Estos modelos son a menudo derivados de la traducción automática estadística [19] y las probabilidades son obtenidas de la observación de pares en función de los parámetros del modelo empleado.

Pero a pesar de la existencia de diversos métodos a nivel de palabra, las tareas de equivalencia y alineación aún están lejos de ser un trabajo trivial debido a la diversidad de idiomas naturales. Por ejemplo, la correspondencia de palabras dentro de las expresiones idiomáticas y traducciones libres son problemáticas. Además, cuando dos idiomas difieren ampliamente en el orden de las palabras, resulta muy difícil encontrar las relaciones. Por consiguiente, es necesario incorporar información lingüística útil para aliviar estos problemas.

## III. Nuestro acercamiento

El propósito de nuestra investigación en una etapa inicial consistió en el diseño de un algoritmo efectivo de alineación de palabras. La idea de obtener equivalencias léxicas en una primera fase, vino con la disponibilidad del recurso MultiWordNet y su concepción.

Se han utilizado al menos dos metodologías de construcción de redes de palabras multilingües. La primera consiste en la construcción independiente de wordnets específicas del lenguaje, con una fase posterior de búsqueda de correspondencias entre ellas. Este enfoque se basa en la hipótesis de que traducciones recíprocas en textos paralelos deben tener los mismos significados y utiliza un índice interlingua (ILI) para materializar las relaciones entre idiomas. EuroWordNet [27] y BalkaNet fueron desarrolladas empleando este enfoque. La segunda metodología consiste en la construcción de las de wordnets específicas del lenguaje manteniendo, tanto como sean posibles, las relaciones semánticas disponibles en Princeton WordNet (PWN). Este acercamiento se empleó para el desarrollo de MultiWordNet [28].

MultiWordNet (MWN) es una base de datos léxica multilingüe, en la cual se ha realizado una alineación estricta entre PWN y redes de palabras para el español y el italiano,

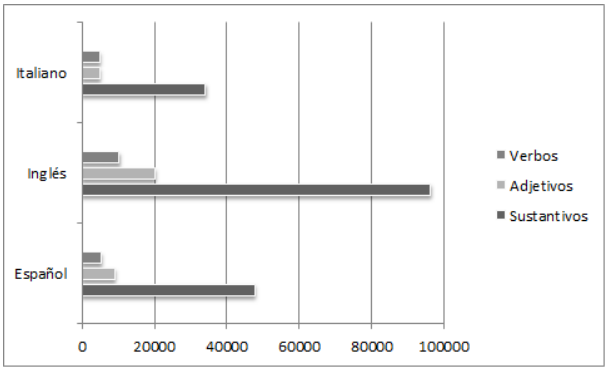



Fig. 1. Composición de las redes de palabras según la categoría gramatical

entre otros lenguajes. Estas redes describen las relaciones léxicas y semánticas existentes entre las palabras y recopilan sus sentidos al igual que un diccionario monolingüe.

La Figura 1 muestra la composición, por categoría gramatical, de cada una de las redes de palabras empleadas en el estudio. Con ello se tiene una perspectiva más clara de la completitud de las mismas, comparando los tres idiomas involucrados.

Los synsets, para cada uno de los idiomas alineados, fueron creados en correspondencia con los synsets de PWN en la medida de las posibilidades. Las relaciones semánticas también fueron importadas de los synsets ingleses correspondientes. Es decir, se asume que si existen dos synsets relacionados en PWN, la misma relación existe en los synsets pertinentes en los otros idiomas [29].

Esta estructura es la que nos permite realizar agrupamiento de sentidos para extraer pares de equivalentes de traducción, mismos que son usados en la fase de alineación, como información lingüística útil para reforzar los enlaces obtenidos. Aunque los experimentos han sido realizados usando MWN, se pueden extender las nociones intuitivas del algoritmo propuesto, introduciendo un marco formal y conceptual para el uso de los ILI. De esta forma, se podrían realizar estudios con redes de palabras construidas bajo el enfoque de EuroWordNet.

### *III-A. Pre-procesamiento de los textos*

Para extraer las equivalencias léxicas, el algoritmo propuesto requiere que los textos se encuentren alineados a nivel de oración. Además, los enlaces pueden producirse entre dos palabras como entradas básicas, dos lemas obtenidos de la aplicación de reglas morfológicas sobre las palabras, o la combinación de éstas. Por ejemplo, el establecimiento de la equivalencia entre: *clothes* (en inglés) y *ropas* (en español) se representa por medio de un enlace entre una palabra y su traducción en plural, es decir, entre una entrada léxica básica y otra a la que se le ha aplicado la regla morfológica de formación del plural. Por tanto, todas las palabras implicadas en ambos contextos deben ser lematizadas.

Como el algoritmo propuesto se basa en la concordancia de synstes en redes de palabras alineadas, una vez que se cuenta con los lemas, se extraen los synsets de las palabras no funcionales. Es importante recalcar que no se lleva a cabo ningún proceso de etiquetado durante esta fase.

### *III-B. Extracción de equivalencias léxicas*

MWN fue elegido como léxico especializado, porque contiene redes de palabras para cada uno de los idiomas involucrados y por su alineación exacta con PWN. Esta última característica constituye la clave en el método que se plantea.

El hecho de que las redes de palabras están alineadas, implica que para expresar un determinado significado, todos utilizan el mismo synset, independiente del idioma. La constitución de los synsets, permite incorporar un grupo de formas sinónimas, por tanto, todos los wordnets en dicho synset almacenan las formas de palabras que pueden representar el significado específico.

Haciendo énfasis en el proceso de extracción de equivalencias, lo que se hace a grandes rasgos, es buscar todos los synsets de la palabra a corresponder, es decir, todos sus posibles sentidos. Este conjunto se compara entonces con los conjuntos extraídos del contexto paralelo (texto en el idioma meta), y aquella palabra que posea mayor coincidencia o intercepción de synsets, será la palabra que se relaciona con la palabra en el origen.

Aquí es necesario considerar los niveles de ambigüedad de los idiomas, pues ello podría afectar en el desempeño del algoritmo. Se realizó entonces el cálculo de los promedios de sentidos por palabra almacenada en su red correspondiente. El análisis comprende la observación por categoría gramatical.

TABLA I
ESTADÍSTICAS DE SENTIDOS EN MWN

| | Español | Italiano | Inglés |
|---|---|---|---|
| Sentidos asignados | 93425 | 64384 | 171018 |
| Sentidos sustantivos | 63028 | 48376 | 119050 |
| Sentidos adjetivos | 17999 | 6228 | 29883 |
| Sentidos verbos | 12398 | 9780 | 22085 |
| Promedio de sentidos | **1.55612** | 1.50009 | 1.42733 |
| Promedio sustantivos | 1.32046 | 1.41351 | 1.23591 |
| Promedio adjetivos | 1.98818 | 1.26972 | 1.48148 |
| Promedio verbos | 2.34057 | 1.99755 | 2.13815 |

La primera fila en la tabla I muestra la cantidad de sentidos que han sido atribuidos a los lemas de entrada en las tablas de índice. De esta forma, si un mismo sentido ha sido asignado a dos palabras diferentes, se toma dos veces en cuenta en este conteo. Por tanto, no se está haciendo referencia al total de synsets del idioma, sino a la cantidad de veces que éstos han sido asignados. Las siguientes tres filas, se corresponden con la división de esta cantidad por POS.

Con la medida de asignación y el número de lemas, se puede determinar el promedio de sentidos. Mientras mayor sea el resultado obtenido, mayor será el grado de polisemia en el lenguaje. Por tanto, el español es, de los tres idiomas

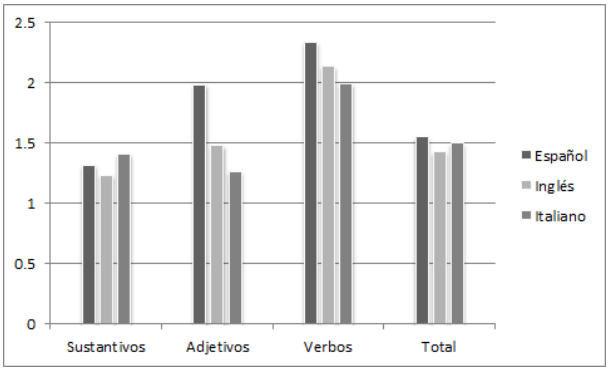



Fig. 2. Promedio de sentidos asociados en MWN

estudiados, el más ambiguo. Sin embargo, en este punto es necesario considerar que aspectos como la incompletidud de los wordnets influyen en los resultados.

En la Figura 2 se presentan los valores de la tabla anterior con formato de gráfica de barras.

Por otra parte, el mecanismo desarrollado está enfocado en la extracción de equivalencias con limitación uno a uno [24]. Es decir, todas las palabras participan en una única equivalencia, con excepción del nulo. Esto para evitar correspondencias no deseadas, como por ejemplo, tener muchas palabras en español alineadas con la misma palabra en inglés. Así, en cada fase del algoritmo, si una traducción potencial pasa a ser parte de un par de equivalencia, la lista de traducciones potenciales se va reduciendo.

En MWN se manejan elementos léxicos simples, por tanto las expresiones multipalabras no son encontradas y consideraremos sólo equivalencias 1:1, durante la fase de extracción. De este modo, en el texto quedan descartadas todas aquellas frases indivisibles con un sentido específico.

### *III-C. El programa de linkeo*

El sistema toma como entrada los textos paralelos preprocesados para realizar el emparejamiento entre las palabras. Para ello, por cada par de sentencias alineadas $< SO_i/SM_j >$, se recorre desde la primera hasta la última palabra en $SO_i$, donde $SO_i$ representa la i-ésima oración en el texto origen y $SM_j$ la j-ésima oración en el texto meta. Si la palabra considerada ($W_x$) pertenece a alguna de las clases gramaticales contempladas en el estudio, es comparada entonces con todas las palabras ($T_x$) en su contexto paralelo.

En la extracción de equivalencias y la alineación a nivel de palabras no existe restricción alguna de la clase gramatical (POS) de las palabras en un par, pues es muy frecuente que durante la traducción se cambie el POS para expresar la misma idea. Por ejemplo, en los fragmentos: *...che non voglio ricordare come si chiami...*(italiano) y *...de cuyo nombre no quiero acordarme...*(español), *nombre* puede ser alineado con *si chiami*.

Basados en el supuesto de que los textos origen y meta se encuentran alineados a nivel de oración, el contexto paralelo de cada palabra en $SO_i$ será aquella oración $SM_j$ en el par alineado. Hasta este momento, todas las palabras no funcionales en $SM_j$ constituyen traducciones potenciales de la palabra $W_x$. La comparación consiste en la búsqueda del conjunto de intersección entre los synsets de los lemas de $W_x$ y los synsets de los lemas de cada palabra en el contexto paralelo, obtenidos en la fase de preprocesamiento. La tabla II representa este análisis. $S()$ y $L()$ representan funciones de extracción de synsets y lemas respectivamente.

TABLA II
BÚSQUEDA DE COINCIDENCIAS DE SYNSETS ENTRE $W_x$ Y LAS PALABRAS DEL TEXTO META

| | $W_x$ |
|---|---|
| $T_1$ | $S(L(W_x)) \cap S(L(T_1))$ |
| $T_2$ | $S(L(W_x)) \cap S(L(T_2))$ |
| ... | ... |
| $T_n$ | $S(L(W_x)) \cap S(L(T_N))$ |

Para determinar la equivalencia léxica de $W_x$ se calcula la cardinalidad de los conjuntos de intersección producidos. A partir de estos valores puede presentarse alguna de las tres siguientes situaciones:

1. Que una de las palabras en el contexto paralelo pueda ser directamente asignada como equivalencia léxica de $W_x$ por ser partícipe del conjunto de intersección de mayor cardinalidad.
2. Que varias palabras en el contexto paralelo poseen la misma cardinalidad.
3. Que ninguna de las palabras en el contexto paralelo constituya la equivalencia de $W_x$

En el primer caso el proceso de extracción de equivalencia queda concluido para la palabra $W_x$.

Cuando no es posible seleccionar una de las traducciones potenciales por haber muchas que poseen el mismo valor de cardinalidad, se utiliza el sentido más cercano. Por ejemplo, suponga que la palabra *quiero* ha tenido un synset en común con *have* y uno con *desire*. Para determinar cuál de las dos traducciones será elegida como equivalente, se compara la posición que ocupa en cada una, el synset coincidente. Aquella palabra cuya posición del synset coincidente sea menor, será la ganadora. La Figura 3 muestra este procedimiento para el ejemplo citado.

Para la forma de palabra *have*, el synset coincidente (v#01530096) ocupa la posición 15, mientras que para *desire*, el synset coincidente (v#01245362) está en la posición 1. Lo anterior se resumen en que es mucho más común utilizar *desire* en el sentido de querer, que *have*.

En el último caso, donde todas las palabras en el contexto paralelo tienen intersección nula con el conjunto de synstes de $W_x$, se han aplicado cuatro medidas de similitud semántica: Leacock and Chodorow [30], Hirst and St-Onge [31], edge [32] y random. Estas medidas han sido implementadas en WordNet::Similarity package [33] y todas ellas se basan, de algún modo, en la estructura de PWN. A continuación se describe cada una de estas cuatro medidas:

$< SO_i$ / $SM_j >$ = <"*En un lugar de la Mancha, de cuyo nombre no* ***quiero*** *acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor.*" /
"*In a village of La Mancha, the name of which I* ***have*** *no* ***desire*** *to call to mind, there lived not long since one of those gentlemen that keep a lance in the lance-rack, an old buckler, a lean hack, and a greyhound for coursing.*">
$W_x$ = *quiero*
posibles traducciones = {*have*, *desire*}
$L(quiero)$ = {$querer$[Verb]}
$L(have)$ = {$have$[Verb]}
$L(desire)$ = {$desire$[Verb]}
$S(querer$[Verb]) = {v#00472243, v#00479719, v#00479841, v#00808096, v#01211759, v#01245362, **v#01530096**}
$S(have$[Verb]) = {v#01508689, v#01794357, v#01443215, v#01509295, v#01857688, v#00080395, v#00786286, v#01620370, v#01185771, v#01509557, v#01876679, v#00080645, v#00045715, v#00523422, **v#01530096**, v#01513366, v#00045966, v#01608899, v#00039991, v#00978092}
$S(desire$[Verb]) = {v#01245362, v#01246466, v#01246175}

Fig. 3. Búsqueda de coincidencias

- *Leacock and Chodorow* (*LCH*): Se basa en la longitud $len(s_1, s_2)$ de la ruta más corta entre dos synsets y la profundidad máxima $D$ de la taxonomía:

$$LCH(s_1, s_2) = \frac{-log\ len(s_1, s_2)}{2D}$$

El valor máximo de esta medida dependerá de la profundidad de la taxonomía y se obtiene al comparar un synset consigo mismo.

- *Hirst-St-Onge* (*HSO*): La idea en esta medida de similitud es que dos conceptos lexicalizados son semánticamente cercanos si sus synsets en PWN están conectados por una ruta que no es tan larga y que no cambia de dirección frecuentemente. Las direcciones de los enlaces en la misma ruta pueden variar (entre: hacia arriba –hiperonimia y meronimia–, hacia abajo –hiponimia y holonimia– y horizontal –antonimia–). La cercanía de la relación está dada por:

$$HSO(s_1, s_2) = C - \text{longituddelaruta} - k \times d,$$

donde $C$ y $k$ son constantes (en la práctica se usa $C$ = 8 y $k$ = 1) y $d$ es el número de cambios de dirección en la ruta. Si no existe dicha ruta, entonces $HSO$ es cero y los synsets se consideran independientes.
- *Edge*: Realiza un conteo de la ruta más corta con respecto a la distancia entre synsets.
- *Random*: Asigna un número aleatorio entre 0 y 1 como medida de similitud, donde 0 indica que no existe similitud entre los conceptos y 1 que poseen el mismo sentido.

Para todas las medidas anteriores, con excepción de Edge, mientras más corta es la ruta entre los synsets, mayor será el valor de similitud.

La utilidad *similarity.pl* permite al usuario introducir pares de conceptos en la forma $word\#pos\#sense$ para medir qué tan parecidos son semánticamente. Por ejemplo, en la tabla III aparecen los sentidos de la palabra *breed* como sustantivo.

TABLA III
SYNSETS ASOCIADOS A CADA SENTIDO DE LA PALABRA *breed* COMO SUSTANTIVO

| Entrada en similarity.pl | Synset correspondiente |
|---|---|
| $breed\#n\#0$ | $n\#06037479$ |
| $breed\#n\#1$ | $n\#06037015$ |
| $breed\#n\#2$ | $n\#07308064$ |
| $breed\#n\#3$ | $n\#03852666$ |

Esta forma de especificar las entradas presenta una desventaja, pues sólo se aceptan palabras inglesas y las posiciones que ocupan los sentidos definidos de dichas palabras en PWN. Es decir, *similarity.pl* no se puede aplicar directamente si se están comparando palabras de idiomas diferentes al inglés.

Para solucionar este problema se obtiene la traducción inglesa para cada uno de los synset de la palabra española o italiana implicada en el par de comparación, aprovechando nuevamente la ventaja de que las redes de palabras en MultiWordNet están alineadas.

Así por ejemplo, si se tiene la palabra *generación*, cuyos synstes como sustantivo son: n#06196326, n#06195881, n#10955750 y n#00546392; las posibles traducciones para dichos synstes serían las palabras inglesas que poseen dichos offsets. Sin embargo, como las posibles traducciones para cada synset se emplean para hacer referencia al mismo concepto (mismo sentido), se puede elegir la primera traducción para cada synset.

TABLA IV
SELECCIÓN DE LA TRADUCCIÓN PARA CADA SYNSET DE LA PALABRA *generación* COMO SUSTANTIVO

| Synset | Posibles traducciones | Traducción elegida |
|---|---|---|
| $n\#06196326$ | *coevals*<br>*contemporaries*<br>*generation* | *coevals* |
| $n\#06195881$ | *generation* | *generation* |
| $n\#10955750$ | *generation* | *generation* |
| $n\#00546392$ | *generation*<br>*multiplication*<br>*propagation* | *generation* |

Ahora sólo bastaría determinar la posición que ocupa el sentido buscado sobre la lista de sentidos de la traducción y se podrían realizar comparaciones entre las palabras

*breed* y *generación* con el formato aceptado por el paquete *similarity.pl*, por ejemplo usando las entradas: breed#n#0 / generation#n#1

Una vez que se tienen los valores de similitud[1] de todas las posibles parejas de synstes que se pueden formar con las dos palabras que se comparan, se toma como similitud del par, el mayor valor obtenido.

| |
|---|
| $coevals\#n\#0$ - $breed\#n\#0$ = 0 |
| $coevals\#n\#0$ - $breed\#n\#1$ = 0 |
| $coevals\#n\#0$ - $breed\#n\#2$ = 2 |
| $coevals\#n\#0$ - $breed\#n\#3$ = 2 |
| $generation\#n\#1$ - $breed\#n\#0$ = 0 |
| $generation\#n\#1$ - $breed\#n\#1$ = 0 |
| $generation\#n\#1$ - $breed\#n\#2$ = 2 |
| $generation\#n\#1$ - $breed\#n\#3$ = 2 |
| $generation\#n\#2$ - $breed\#n\#0$ = 0 |
| $generation\#n\#2$ - $breed\#n\#1$ = 0 |
| $generation\#n\#2$ - $breed\#n\#2$ = 4 |
| $generation\#n\#2$ - $breed\#n\#3$ = 0 |
| $generation\#n\#3$ - $breed\#n\#0$ = 0 |
| $generation\#n\#3$ - $breed\#n\#1$ = 0 |
| $generation\#n\#3$ - $breed\#n\#2$ = 0 |
| $generation\#n\#3$ - $breed\#n\#3$ = 0 |

Fig. 4. Similitud de *breed* y las traducciones de *generación* (*coevals* y *generation*)

Este valor se coloca en una matriz de similitud como se muestra en la Figura 5. Tal matriz posee el valor $-1$ en todas las palabras que fueron directamente asignadas por poseer la mayor cardinalidad absoluta, con el objetivo de que no sean tomadas nuevamente en cuenta en la actual etapa de extracción. Así, de la matriz se advierte que $W_1$ y $T_2$ fueron previamente relacionadas en un par de equivalencia.

| | $T_1$ | $T_2$ | *breed* | ... | $T_n$ |
|---|---|---|---|---|---|
| $W_1$ | $-1$ | $-1$ | $-1$ | ... | $-1$ |
| *generación* | 2 | $-1$ | 4 | ... | 0 |
| $W_3$ | 0 | $-1$ | 3 | ... | 2 |
| ... | ... | ... | ... | ... | ... |
| $W_m$ | 3 | $-1$ | 2 | ... | 0 |

Fig. 5. Matriz de similitud

Finalmente, los pares de equivalencia se forman comenzando por la mayor similitud y así sucesivamente. En el caso de la matriz anterior los pares que se obtendrían con los valores definidos son:

$$(generación, breed)$$

$$(W_3, T_n)$$

$$(W_m, T_1)$$

[1]En este ejemplo obtenidos empleando el método Hirst-St-Onge

## IV. Evaluación

Para realizar nuestros experimentos, empleamos fragmentos elegidos aleatoriamente de la novela *Don Quijote de la Mancha*, en sus versiones paralelas español / inglés y español / italiano. Los fragmentos usados en ambos pares de idiomas fueron los mismos. Los conjuntos de prueba estuvieron formados por 23 sentencias alineadas. El texto en español está constituido por 828 palabras. La tabla V muestra la composición de cada uno de los fragmentos de los textos meta empleados como corpus de prueba.

TABLA V
Composición de los textos meta

| | Inglés | Italiano |
|---|---|---|
| # palabras | 796 | 866 |
| Promedio de sentidos | 6.8653 | 4.4276 |

Para producir el gold estándar de los pares de alineación de todas las palabras no funcionales, dos anotadores fueron instruidos con procedimientos específicos de cuándo asignar un equivalente nulo. No se incluyeron etiquetas de probabilidades. En caso de que hubiera un desacuerdo para un par específico, un tercer anotador definía el correcto.

La tabla VI muestra la cantidad de pares de alineación determinados por los anotadores (estándar de oro). El topline indica el número máximo de equivalencias que podrían ser extraídas por el sistema, considerando relaciones 1:1 exclusivamente, la incompletitud de las redes de palabras que conforman MWN y las brechas léxicas (gaps) de los lenguajes implicados.

TABLA VI
Número de equivalencias sugeridas por los anotadores y topline / recall del sistema

| Texto | Gold estándar | Topline |
|---|---|---|
| Don Quijote de la Mancha (español / inglés) | 389 | 329 |
| Don Quijote de la Mancha (español / italiano) | 333 | 277 |

### *IV-A. Medidas de evaluación*

Realizamos la evaluación respecto a tres diferentes medidas: precisión, recall y $F$-measure. La precisión es calculada como el número de equivalencias extraídas correctamente entre el número de equivalencias sugeridas por el sistema. El recall se corresponde al número de equivalencias extraídas correctamente entre el número de equivalencias sugeridas por los anotadores.

Sin embargo, ni la precisión ni el recall pueden, de manera independiente, determinar la calidad del emparejamiento. Por lo general, la maximización del recall compromete la precisión y viceversa [34]. Por tanto, se requiere una medida que combine ambos parámetros. La $F$-measure posee esta característica y se determina como:

$$F\text{-measure} = \frac{2 \times \text{precisión} \times \text{recall}}{\text{precisión} + \text{recall}}$$

En este caso, la precisión y el recall poseen el mismo peso, pero la fórmula anterior puede ajustarse si se desea otorgar mayor peso a algunas de estas dos medidas.

### *IV-B. Resultados y discusión*

Los resultados fueron obtenidos con una iteración del algoritmo, por tanto, los tiempos de respuesta son muy pequeños. Además, se establecieron umbrales para la asignación de similitudes en cada método. De esta forma, se evitan despropósitos lingüísticos, al establecer qué valores numéricos de las medidas se consideran cercanos. Los umbrales en cada método fueron los siguientes: LCH $\Rightarrow$ 2, HSO $\Rightarrow$ 2, Edge $\Rightarrow$ 0.2 y Random $\Rightarrow$ 0.2. De esta forma, un valor de 0.166667 para el par (*astillero*, *village*) en el método Edge es descartado y se asigna similitud 0, o sea, sólo se toman valores de similitud mayores a 0.2 como fue señalado en el umbral. Esto evita la extracción de equivalencias erróneas y por tanto, tiene influencia en la precisión.

La tabla VII muestra la cantidad de pares extraídos por el sistema (total), la cantidad de éstos que son correctos (considerando equivalencia –EQ– y alineación –AL–) y los valores de las tres medidas empleadas. Los mismos valores han sido graficados en la Figura 6. La entrada *Sólo coincidentes*, se refiere al caso donde no se ha aplicado ninguna medida de similitud semántica para efectuar la extracción cuando no hay synsets comunes entre la palabra a corresponder y sus traducciones potenciales.

Si se comparan los valores de las tres medidas, independientemente del método de similitud empleado, existe una diferencia aproximada de 3.76 % en promedio entre el procedimiento de extracción de equivalencias (EQ) y la alineación (AL). Es comprensible que en EQ siempre se obtengan mejores resultados, pues, por definición, un par se considera correcto al existir al menos un contexto en el que las palabras que conforman el par sean traducciones entre sí.

El desempeño del sistema se ve afectado durante la alineación, ya que no sólo se deben considerar pares de traducción, sino también la correspondencia entre las palabras según la función que realizan en la oración y en este proceso, la inserción de nulos es requerida.

En términos generales, se puede advertir que el inglés consigue mejores resultados, en su alineación con el español, para todos los métodos aplicados, a pesar de tener un mayor grado de polisemia para el corpus de prueba (véase la tabla V).

La precisión del método de "Sólo coincidentes" tiene que ver con la asertividad de las alineaciones de las redes de palabras que conforman MWN. Sin embargo, los valores de recall obtenidos para este método son pobres en ambos casos (usando el gold standard y el topline). Esto se debe a dos razones fundamentales: 1) el hecho de que en MWN se almacenan elementos léxicos simples y por tanto, es imposible asignar un sentido específico a las expresiones multipalabra y 2) la incompletitud de las redes (véase Figura 1 para observar la desproporción entre el inglés y el resto de los idiomas).

Por otra parte, la medida LCH es comparable en términos de $F$-measure con el método de sólo coincidencia, a pesar de la diferencia significativa si se analizan los valores de precisión. Los métodos de similitud HSO y Edge, aumentan el recall, aunque no de manera significativa (entre 2 y 6 % aproximadamente), pero lo hacen a costa de una disminución considerable de la precisión (entre 16 y 18 %). En este sentido, es lógico que el método HSO posea el mayor valor de recall, puesto que su definición toma en cuenta cuatro tipos de relaciones semánticas (hiperonimia, meronimia, hiponimia y holonimia) y una léxica (antonimia) de PWN, en tanto que LCH sólo se basa en la hiperonimia.

La baja precisión del método Random, con respecto al resto de las medidas, está relacionada con el simple marcaje basado en la aleatoriedad de la asignación del valor de similitud.

El comportamiento anterior, es también notorio si se toman los valores del topline en vez de los del gold estándar (descartando las equivalencias establecidas por pertenencia a frases, incompletitud de MWN y gaps), como se observa en la tabla VIII.

La Figura 7 muestra la mejora en los valores de $F$-measure, comparando los resultados obtenidos con el topline y el gold estándar.

Si se colocan en el mismo gráfico la precisión y el recall, como se muestra en la Figura 8, se puede advertir que la precisión disminuye a medida que aumenta el recall, independientemente de la base (gold standard o topline). Para cuantificar la relación que existe entre estas medidas se ha determinado el coeficiente de correlación producto o momento de Pearson, $r$, un índice adimensional acotado entre $-1$ y 1 que refleja el grado de dependencia lineal entre dos conjuntos de datos.

$$r = \frac{\sum_{i=1}^{n}(x-\overline{x})^2(y-\overline{y})^2}{\sqrt{\sum_{i=1}^{n}(x-\overline{x})^2 \sum_{i=1}^{n}(y-\overline{y})^2}}$$

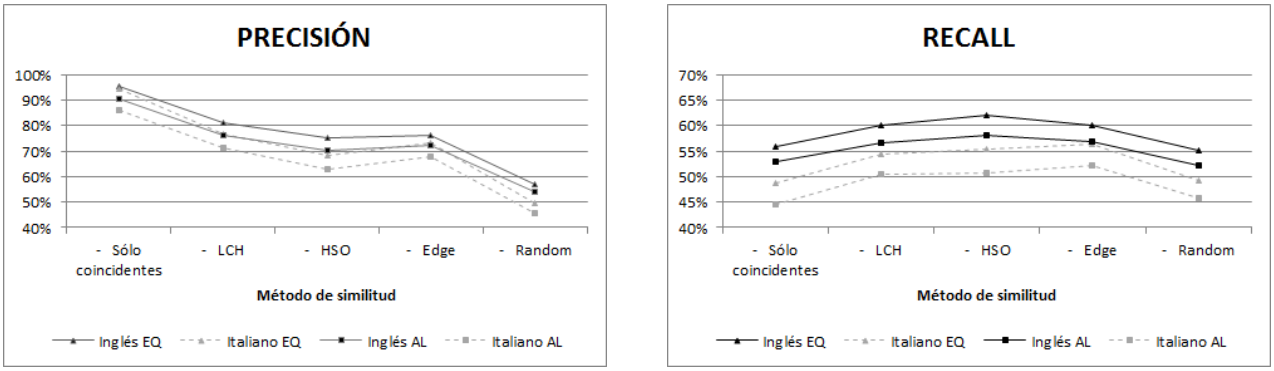
donde:

- $n$ es el número de métodos de similitud empleados
- $x$ el valor de precisión para el método $i$ (conjunto de valores independientes)
- $y$ el valor de recall para el método $i$ (conjunto de valores dependientes)
- $\overline{x}$ e $\overline{y}$ son las medias de muestra promedio para el conjunto de valores independientes y dependientes respectivamente.

El resultado de aplicar el coeficiente de correlación a los valores de precisión y recall obtenidos en la tabla VII para la equivalencia son $r = -0{,}9675$ (inglés) y $r = -0{,}9497$ (italiano) y para la alineación $r = -0{,}9838$ (inglés) y $r =$

TABLA VII
EVALUACIÓN DE LOS RESULTADOS EN EL CORPUS DE PRUEBA

| Versión (Gold-estándar) | Método similitud | Total | Correctos EQ | Correctos AL | Precisión EQ | Precisión AL | Recall EQ | Recall AL | $F$-measure EQ | $F$-measure AL |
|---|---|---|---|---|---|---|---|---|---|---|
| español/inglés (389) | Sólo coincidentes | 228 | 217 | 206 | 95.18 % | 90.35 % | 55.78 % | 52.96 % | 70.34 % | 66.78 % |
| | LCH | 289 | 234 | 220 | 80.97 % | 76.12 % | 60.15 % | 56.56 % | 68.94 % | 64.88 % |
| | HSO | 321 | 241 | 226 | 75.08 % | 70.40 % | 61.95 % | 58.10 % | 67.89 % | 63.66 % |
| | Edge | 307 | 234 | 221 | 76.22 % | 71.99 % | 60.15 % | 56.81 % | 67.24 % | 63.51 % |
| | Random | 377 | 214 | 203 | 56.76 % | 53.85 % | 55.01 % | 52.19 % | 55.87 % | 53.01 % |
| español/italiano (333) | Sólo coincidentes | 172 | 162 | 148 | 94.19 % | 86.05 % | 48.65 % | 44.44 % | 64.16 % | 58.61 % |
| | LCH | 236 | 181 | 168 | 76.69 % | 71.19 % | 54.35 % | 50.45 % | 63.62 % | 59.05 % |
| | HSO | 269 | 184 | 169 | 68.40 % | 62.83 % | 55.26 % | 50.75 % | 61.13 % | 56.15 % |
| | Edge | 257 | 188 | 174 | 73.15 % | 67.70 % | 56.46 % | 52.25 % | 63.73 % | 58.98 % |
| | Random | 332 | 164 | 152 | 49.40 % | 45.78 % | 49.25 % | 45.65 % | 49.32 % | 45.71 % |




(a) (b)

Fig. 6. Valores de precisión y recall por versión, procedimiento y método de similitud

TABLA VIII
EVALUACIÓN DE LOS RESULTADOS EN EL CORPUS DE PRUEBA

| Versión (Topline) | Método similitud | Total | Correctos EQ | Correctos AL | Precisión EQ | Precisión AL | Recall EQ | Recall AL | $F$-measure EQ | $F$-measure AL |
|---|---|---|---|---|---|---|---|---|---|---|
| español/inglés (329) | Sólo coincidentes | 221 | 217 | 206 | 98.19 % | 93.21 % | 65.96 % | 62.61 % | 78.91 % | 73.96 % |
| | LCH | 271 | 234 | 220 | 86.35 % | 81.18 % | 71.12 % | 66.87 % | 78.00 % | 73.33 % |
| | HSO | 299 | 241 | 226 | 80.60 % | 75.59 % | 73.25 % | 68.69 % | 76.75 % | 71.98 % |
| | Edge | 288 | 234 | 221 | 81.25 % | 76.74 % | 71.12 % | 67.17 % | 75.85 % | 71.64 % |
| | Random | 346 | 214 | 203 | 61.85 % | 58.67 % | 65.05 % | 61.70 % | 63.41 % | 60.15 % |
| español/italiano (277) | Sólo coincidentes | 165 | 162 | 148 | 98.18 % | 89.70 % | 58.48 % | 53.43 % | 73.30 % | 66.97 % |
| | LCH | 216 | 181 | 168 | 83.80 % | 77.77 % | 65.34 % | 60.65 % | 73.43 % | 68.15 % |
| | HSO | 236 | 184 | 169 | 77.97 % | 71.61 % | 66.43 % | 61.01 % | 71.74 % | 65.89 % |
| | Edge | 232 | 188 | 174 | 81.03 % | 75.00 % | 67.87 % | 62.82 % | 73.87 % | 68.37 % |
| | Random | 291 | 164 | 152 | 56.36 % | 52.23 % | 59.21 % | 54.87 % | 57.75 % | 53.52 % |

$-0{,}9194$ (italiano), lo que indica que estas medidas poseen un alto vínculo y son inversamente dependientes.

## REFERENCIAS


[1] E. Macklovitch and M. Hannan, "Line 'em up: Advances in alignment technology and their impact on translation support tools," *Machine Translation*, vol. 13, no. 1, pp. 41–57, 1998.

[2] C. . Nevill and T. Bell, "Compression of parallel texts," *Information Processing and Management: an International Journal*, vol. 28, no. 6, pp. 781–793, 1992.

[3] J. Vera and G. Sidorov, "Proyecto de preparación del corpus paralelo alineado español-inglés," in *Memorias del Encuentro Internacional de la Ciencias de la Computación*, Mexico, 2004.

[4] D. Tufiş, A. Barbu, and R. Ion, "Treq-al: a word alignment system with limited language resources," in *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond*, Canada, 2003, pp. 36–39.

[5] R. Mihalcea and T. Pedersen, "An evaluation exercise for word alignment," in *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Canada, 2003, pp. 1–10.

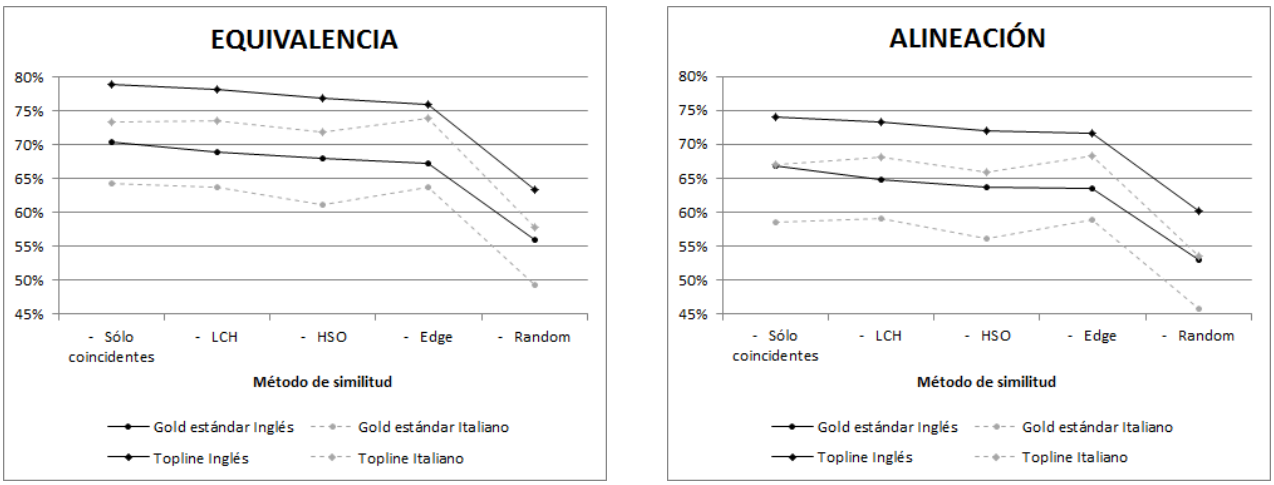
[6] C. Kit, J. Webster, H. P. K. Sin, and H. Li, "Clause alignment for bilingual hong kong legal texts with available lexical resources," in *Proceedings of the 20th International Conference on Computer Processing of Oriental Languages*, China, 2003, pp. 286–292.

[7] W. Gale and K. Church, "Program for aligning sentences in bilingual corpora," *Computational Linguistics*, vol. 19, no. 1, pp. 75–102, 1993.

[8] P. Brown, J. Lai, and R. Mercer, "Aligning sentences in parallel corpora," in *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, EUA, 1991.

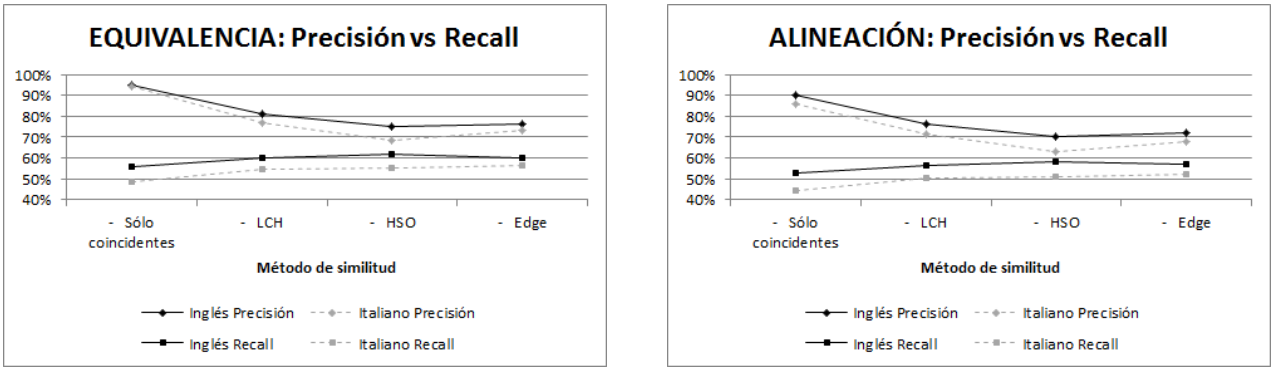
[9] D. Wu, "Aligning a parallel english-chinese corpus statistically with lexical criteria," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, EUA, 1994, pp. 80–87.

[10] A. Gelbukh, G. Sidorov, and J. Vera, "A bilingual corpus of novels aligned at paragraph level," in *Proceedings of the 5th International*




(a) (b)

Fig. 7. Comparación de los valores de $F$-measure con base en el gold estándar y el topline




(a) (b)

Fig. 8. Comparación de los valores de precisión y recall con base en el gold estándar

*Conference on NLP*, Finlandia, 2006, pp. 16–23.

[11] M. Kay and M. Roscheisen, "Text-translation alignment," *Computational Linguistics*, vol. 19, no. 1, pp. 121–142, 1993.

[12] S. Chen, "Aligning sentences in bilingual corpora using lexical information," in *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, EUA, 1993, pp. 9–16.

[13] M. Haruno and T. Yamazaki, "High-performance bilingual text alignment using statistical and dictionary information," in *Proceedings of the Annual Conference of the Association for Computational Linguistics*, EUA, 1996, pp. 131–138.

[14] M. Mikhailov, "Parallel corpus aligning: Illusions and perspectives," *The Austrian Academy Corpus*, 2002.

[15] T. Tanaka and Y. Matsuo, "Extraction of translation equivalents from non-parallel corpora," in *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation*, England, 1999, pp. 109–119.

[16] F. Smadja, V. Hatzivassiloglou, and K. McKeown, "Translating collocations for bilingual lexicons: A statistical approach," *Computational Linguistics*, vol. 22, no. 1, pp. 1–38, 1996.

[17] M. Simard, G. Foster, and P. Isabelle, "Using cognates to align sentences in parallel corpora," in *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Canada, 1992, pp. 67–81.

[18] F. Debili and E. Sammouda, "Appariement des phrases de textes bilingues français-anglais et français-arabe," in *Proceedings of the 14th Conference on Computational Linguistics*, Francia, 1992.

[19] P. Brown, S. Della, V. Della, and R. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[20] S. Vogel, H. . Ney, and C. Tillmann, "Hmm-based word alignment in statistical translation," in *Proceedings of the 16th International Conference on Computational Linguistics*, Denmark, 1996, pp. 836–841.

[21] K. Sato and H. Saito, "Extracting word sequence correspondences with support vector machines," in *Proceedings of the 19th international conference on Computational linguistics*, Taiwan, 2002, pp. 1–7.

[22] D. Tufis, "A cheap and fast way to build useful translation lexicons," in *Proceedings of the 19th International Conference on Computational Linguistics*, Taiwan, 2002, pp. 1030–1036.

[23] S. Ker and J. Chang, "A class-based approach to word alignment," *Computational Linguistics*, vol. 23, no. 2, pp. 313–343, 1997.

[24] D. Melamed, "Models of translational equivalence among words," *Computational Linguistics*, vol. 26, no. 2, pp. 221–249, 2000.

[25] D. Hiemstra, "Deriving a bilingual lexicon for cross language information retrieval," in *Proceedings of Gronics*, Netherlands, 1997, pp. 21–26.

[26] J. Kupiec, "An algorithm for finding noun phrase correspondences in bilingual corpora," in *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, EUA, 1993, pp. 17–22.

[27] "Eurowordnet," http://www.illc.uva.nl/EuroWordNet/, 2001, consultado 29/12/08.

[28] "Multiwordnet," http://multiwordnet.itc.it/english/home.php, 2004, consultado 29/12/08.

[29] E. Pianta, L. Bentivogli, and C. Girardi, "Multiwordnet: developing an aligned multilingual database," in *Proceedings of the First International Conference on Global WordNet*, India, 2002, pp. 21–25.
[30] C. Leacock and M. Chodorow, "Combining local context andwordnet similarity for word sense identification," *WordNet: An electronic Lexical Database*, pp. 265–283, 1998.
[31] G. Hirst and D. St-Onge, "Lexical chains as representations of context for the detection and correction of malapropisms," *WordNet: An electronic Lexical Database*, pp. 305–332, 1998.
[32] R. Rada, H. Mili, E. Bicknell, and M. Bletner, "Development and application of a metric on semantic nets," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 1, pp. 17–30, 1989.
[33] T. Pedersen, S. Patwardhan, and J. Michelizzi, "Wordnet::similarity - measuring the relatedness of concepts," in *Proceedings of the 19th National Conference on Artificial Intelligence*, EUA, 2004, pp. 144–152.
[34] S. M. H.H. Do and E. Rahm, "Comparison of schema matching evaluations," in *Proceedings of the GI-Workshop Web and Databases*, Erfurt, 2002, pp. 221–237.