Clasificación de los métodos para la mejora de WSD y de los métodos de evaluación correspondientes

Alisa Zhila, Alexander Gelbukh

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

Resumen. Se presenta una revisión y comparación de los métodos para la evaluación de la desambiguación de los sentidos de las palabras (WSD, por sus siglas en inglés: *Word Sense Disambiguation*). Basándose en esta revisión, se propone una clasificación características de los modelos de WSD que se pueden modificar para mejorar el modelo.

Palabras clave: precisión, recall, inventario de sentidos, traducción automática, procesamiento del lenguaje natural.

1 Introducción

Las palabras en el lenguaje pueden tener varias sentidos, o acepciones: un banco puede ser un mueble, la orilla del río, un conjunto de peses, o una organización financiera. Tales palabras se llaman *polisémicas*.

En el procesamiento del lenguaje natural (PLN) incluye la tarea de asignar una acepción seleccionada de un inventario de sentidos predeterminado (por ejemplo, un diccionario), a una palabra polisémica en el texto dado. Esta tarea se llama la desambiguación de los sentidos de las palabras (WSD, por sus siglas en inglés: *Word Sense Disambiguation*). Es un proceso muy importante para los sistemas de PLN que realizan el procesamiento del lenguaje a nivel semántico [29]. Sin embargo, es una tarea muy difícil. Aunque los primeros intentos de solucionar este problema se remontan al 1949 [17], la calidad de los métodos actuales no es suficiente [16].

El desempeño de un sistema de WSD se caracteriza por su *cobertura* (*coverage*) C, *precisión* P y *recuerdo* (*recall*) R, los cuales se combinan en la medida cumulativa F:

$$C = \frac{A}{N}, \quad P = \frac{K}{A}, \quad R = \frac{K}{N}, \quad F = \frac{2PR}{P+R},$$

donde N es el número total de las palabras en la entrada, A es el número de las palabras para las cuales el sistema dio alguna respuesta —dado que para ciertas palabras el sistema no propuso ningún sentido: dijo "no sé"— y K es el número de las palabras para las cuales el esta respuesta resultó correcta.

Gale *et al*. [13] estimaron los límites superior e inferior para el desempeño que pueden alcanzar los sistemas de WSD. El límite superior es el grado de acuerdo de anotadores humanos, que es 90%. El límite inferior corresponde al sentido más frecuente: es la proporción de las veces cuando la acepción más frecuente de una palabra es la correcta; es usualmente alrededor del 60% [12].

Hasta el principio de la década de 2000 la WSD fue tratada sobre todo como una tarea aislada, no integrada en alguna aplicación real del PLN. Además, el diccionario WordNet en inglés [11] fue utilizado como el único inventario de sentidos en el área, lo que efectivamente lo convirtió en un inventario de sentidos estándar. Por lo tanto, prácticamente todos los sistemas de WSD se evaluaron en las mismas condiciones, lo que facilitó la comparación del desempeño de los sistemas.

Sin embargo, a partir de mediados de la década de 2000 numerosos autores [1, 2, 16] demostraron que el uso del WordNet como el único inventario de sentidos estándar y la consideración de la WSD como la tarea aislada dirigieron a las soluciones del problema de WSD que se aplicaban mal en las tareas reales del PLN.

Para mejorar el desempeño de los sistemas de WSD, se han sugerido varias medidas: unir las acepciones del WordNet para obtener inventarios de sentidos de granularidad más gruesa [1, 15]; usar las características léxicas de varios idiomas y corpus paralelo [3–6, 18, 19] (este enfoque se conoce como la WSD en varios idiomas o la WSD multilingüe); tratar un problema de WSD como una tarea de traducción de palabras para poder incorporar los módulos de WSD directamente en los sistemas de traducción automática (TA) [2, 8, 9, 10, 27].

Con eso, todas estas modificaciones llevaron a la aparición de ciertas dificultades en la evaluación y comparación del desempeño de los sistemas de WSD, ya que los resultados no eran evaluados con respecto al mismo inventario de sentido estándar.

En este trabajo presentamos nuestra clasificación de las distintas direcciones para la mejora de los métodos para la WSD. A cada una, le corresponde su tipo de evaluación, lo que nos da una clasificación de los métodos para la evaluación de WSD. En las tres siguientes secciones (secciones 2, 3 y 4) analizamos tres diferentes tipos de los métodos o las aplicaciones de WSD, con el propósito de demostrar que en todos ellos se observan las mismas direcciones de mejora y las mismas metodologías de su evaluación. Eso nos lleva a la clasificación que buscamos, la cual la presentamos en la sección 5.

2 Caso de mejoras de la WSD basadas en las modificaciones de los inventarios de sentidos

Uno de los factores que dificultan esta tarea es el hecho de que las acepciones en los diccionarios son discretas, mientras que el significado de una palabra ocupa un diapasón continuo [7, 14].

Navigli [15] sostiene que la granularidad de sentidos demasiada fina es uno de los principales obstáculos para el éxito de la WSD. Él presenta un método para reducir el nivel de granularidad del inventario de sentidos WordNet a través del mapeo de las acepciones del WordNet sobre las acepciones de un diccionario con jerarquía semántica elaborada manualmente, el Diccionario Oxford de Inglés.

Como resultado de este mapeo, Navigli obtiene un inventario de sentidos de granularidad gruesa, donde cada acepción corresponde al menos a un sentido del WordNet. La reducción del número de las acepciones resulta el 33,54% (de 60.302 a 40.079 acepciones) y la disminución del grado de polisemia es de 3,14 a 2,09.

Para evaluar el aumento del desempeño de la WSD con el inventario de sentidos de granularidad más gruesa, Navigli ejecuta cinco sistemas de WSD en la tarea

Allwords en inglés de Senseval-3 [20] con el inventario de sentidos nuevo en lugar de WordNet. De esta manera el autor evalúa el aumento del desempeño de los sistemas de WSD elegidos en comparación contra su desempeño con el inventario de sentidos original de granularidad fina.

La tarea Allwords en inglés de Senseval-3 requiere de los sistemas de WSD asignar una acepción de un inventario de sentidos a 2.081 palabras de un conjunto de 301 oraciones de la ficción, noticias y editoriales.

Los sistemas elegidos fueron los tres mejores sistemas de WSD de Senseval-3: GAMBL [21], SenseLearner [22], y KOC [23], y el mejor sistema sin supervisión IRST-DDD [24]. Navigli también incluye un algoritmo de WSD basado en conocimiento que se llama *Structural Semantic Interconections* [25], ya que éste supera a todos los sistemas que funcionen sin entrenamiento.

Todos los sistemas mostraron un aumento de la medida-F de más de 10% en comparación con sus resultados para la misma tarea con el inventario de sentidos de granularidad fina. El desempeño para la línea base del sentido más frecuente, así como la línea base del sentido al azar también aumentó el 15%.

Snow *et al*. [1] nota que "diferentes aplicaciones de PLN requieren diferentes granularidades de acepciones a fin de aprovechar mejor las diferencias de los sentidos de la palabra, y que para muchas aplicaciones acepciones del WordNet son demasiado finas." En [1] la tarea de la fusión de acepciones se presenta como un problema de aprendizaje supervisado. Como ese tipo de problemas requiere los datos de entrenamiento, los agrupamientos (o los clusters) de acepciones hechos y etiquetados manualmente se usan para entrenamiento. Los autores entrenan un clasificador discriminativo sobre la amplia variedad de características derivadas de la estructura del WordNet, la evidencia basada en corpus, y la evidencia de otros recursos léxicos. Así se obtienen varios WordNets con acepciones agrupadas de granularidad diferente.

Para evaluar el desempeño de la WSD con los inventarios de sentidos de granularidad más gruesa Snow *et al*. aplican un método similar a él de Navigli [15]: "Sabiendo las respuestas anteriores de los sistemas en Allwords en inglés de Senseval-3, podemos evaluar los mismos sistemas en el mismo corpus, pero utilizando el inventario de sentidos de granularidad gruesa proporcionado por nuestra taxonomía de acepciones agrupadas". Como prueba de los sistemas se eligieron los 3 sistemas mejores que participaron en Allwords en inglés de Senseval-3: [21, 22, 23].

Snow *et al*. afirman que es obvio que cualquier agrupación de acepciones aumenta el desempeño de los sistemas de WSD. Por lo tanto, en [1] se introduce un inventario de sentidos aleatorio con la agrupación de las acepciones de la misma granularidad que las obtenidas y se realiza la evaluación de cada sistema para estos tres inventarios de sentidos, a diferencia de la comparación hecha por Navigli, donde se compara el desempeño únicamente con los dos inventarios (fino y grueso) de sentidos.

En [1] se muestra que, aunque la WSD con el inventario de sentidos de granularidad más gruesa supera a la WSD con uno de granularidad fina más del 10% de la medida-F, un máximo de sólo el 3,55% (un promedio por los 3 algoritmos de WSD) de mejora de la medida-F se obtiene si se compara contra el inventario de sentidos aleatorio de la misma granularidad. También agregan que este aumento se puede observar sólo con un umbral adecuado para el agrupamiento aleatorio.

La revisión de los artículos sobre el aumento del desempeño de los sistemas de WSD aislados debido al uso de los inventarios de sentidos de granularidad gruesa, muestra que en ese caso para evaluar el desempeño de la WSD se necesitan:

- varios sistemas de WSD existentes,
- un conjunto de datos estándar para el experimento, por ejemplo una tarea Allwords en inglés del Senseval-3,
- los inventarios de sentidos con diferentes granularidades, el de granularidad gruesa siendo derivado del de granularidad fina,
- preferiblemente, un inventario de sentidos aleatorio con la misma granularidad que la del modificado.

Entonces se compara el desempeño de los sistemas de WSD para el mismo conjunto de datos contra diferentes inventarios de sentidos y los resultados de la comparación se presentan como la evaluación de la mejora introducida en un trabajo.

Admitimos que la cantidad de los trabajos observados en esta sección es bastante limitada. Sin embargo, son precisamente los trabajos que demuestran el enfoque más común para la evaluación de resultados en este tipo de investigaciones.

3 Caso de la WSD multilingüe

Una estrecha relación entre la traducción automática (TA) y la WSD puede ser observada a través de dos ramas de los trabajos: textos paralelos mejorando una tarea de la WSD aislada y la WSD mejorando los existentes sistemas de TA. En esta sección analizamos el primer grupo.

Diab [3] analiza la posibilidad de la creación sin supervisión de un corpus etiquetado con sentidos (o acepciones) para las lenguas que carecen de los recursos existentes y su aplicación para WSD supervisada. En su trabajo la autora se basa en el hecho que lexicalización de un concepto, una noción, entre lenguajes tiende a ser consistente, conservando algunos rasgos fundamentales de su semántica, y al mismo tiempo variable a través de las preferencias de traductor, de los contextos y estilos. Diab extrae las características léxicas de los corpus paralelos y las utiliza para aumentar el desempeño de un sistema de WSD supervisado. Para evaluar sus resultados, la autora compara el desempeño de su sistema mejorado con otros sistemas que participaron en el taller de Senseval-2, ejecutándolos en el conjunto de datos estándar Allwords en inglés de Senseval-2.

Ng *et al*. [4], así como Diab, tienen como objetivo reducir el costo de la anotación de los corpus para la WSD supervisada. Sin embargo, su enfoque es algo diferente a lo de Diab. Ellos usan un corpus paralelo de chino-inglés alineado por palabras de forma automática para etiquetar con las correspondientes traducciones un corpus en inglés que se utiliza para el entrenamiento de un sistema de WSD supervisado. En el caso de los sustantivos elegidos para el etiquetado a menudo pasa la fusión de los sentidos que se consideran como acepciones diferentes en el WordNet inglés con la base de la misma traducción en chino.

Se debe notar que este enfoque es opuesto al de Diab ya que ella usa un conjunto de diferentes pero sinónimas traducciones de una palabra polisémica como las características adicionales para la mejor desambiguación de los sentidos de las

palabras. Sorprendentemente, el enfoque de Ng *et al*. es similar a los descritos en la sección 2 debido a la fusión de sentidos.

Por lo tanto, la evaluación de los resultados en [4] se realizó de acuerdo a la metodología de la sección 2: los autores compararon los resultados para los diferentes etiquetados con sentidos –uno con respecto al WordNet inglés original y el otro con respecto a su inventario de sentidos modificado– para un elegido sistema de WSD [26] en un subconjunto estándar de los sustantivos más difíciles de la tarea de la muestra léxica (en inglés: *lexical sample*) inglesa del Senseval-2.

Ide [5] realiza una investigación dirigida a determinar el grado en que los equivalentes de traducción para los diferentes significados de una palabra polisémica inglesa son lexicalizados diferentemente a través de una variedad de idiomas. También determina si esta información puede utilizarse para estructurar o crear un conjunto de distinciones de sentidos útil para las aplicaciones del PLN. Lamentablemente, en este trabajo la autora no investiga la aplicación real de sus resultados a un sistema de WSD, por lo tanto, no se propone una metodología de evaluación de la WSD.

Sin embargo, en los trabajos posteriores Ide *et al*. [18, 19] proporcionan una investigación más profunda de los resultados obtenidos en [5]. Su investigación es similar a [4], ya que los autores explotan las traducciones obtenidos de corpus paralelos para determinar las diferencias en sentidos que pueden utilizarse para el etiquetado de sentidos automático y otras tareas de la desambiguación. En [5] los agrupamientos de las traducciones correspondientes se utilizaron como etiquetas de sentidos para un conjunto de sustantivos polisémicos y compusieron un inventario de sentidos alternativo. El mismo algoritmo del agrupamiento se usó como el algoritmo de WSD. Para el mismo conjunto de sustantivos polisémicos dos anotadores realizaron etiquetado de sentidos manual como un etiquetado de referencia.

Para evaluar los resultados del etiquetado obtenido automáticamente con el algoritmo de WSD elegido, las etiquetas se compararon de forma manual con las proporcionadas por los anotadores humanos para el mismo conjunto de sustantivos. Aunque el porcentaje total de discordancia entre las etiquetas de los sentidos obtenidas automáticamente con las establecidas por los seres humanos es más del 35%, la diferencia entre cada anotador humano y el algoritmo es sólo el 10-13%.

En esta sección para evaluación de los métodos de WSD se han presentado:

- los métodos de WSD referenciales incluyendo el etiquetado manual,
- los conjuntos de datos experimentales estándares, por ejemplo conjuntos de sustantivos polisémicos ,
- en el trabajo de Ng *et al*. [4] se usaron diferentes inventarios de sentidos.

En resumen podemos concluir la metodología de la evaluación del desempeño de la WSD depende de las modificaciones concretas realizadas en un trabajo y o depende tanto del tipo de la WSD.

4 Caso de la WSD para la traducción automática

Aunque las tareas de la WSD en varios idiomas y la WSD para la TA pueden parecer similares, la diferencia esencial entre estos dos grupos de los enfoques es que

el primero se centra en el uso de recursos multilingües, que normalmente se utilizan en los sistemas de la traducción automática estadística (SMT, por sus siglas en inglés *Statistical Machine Translation*), para mejorar el desempeño de la WSD, mientras que el segundo se centra en el objetivo de utilizar los modelos de WSD para mejorar la calidad de la traducción.

Vickery *et al*. [2] fueron unos de los primeros en centrarse en el diseño de los sistemas de WSD para el propósito específico de la traducción. En [2] se utilizan los métodos de WSD supervisados y el corpus de los textos paralelos alineados por palabras para resolver la tarea de traducción de una palabra. Luego los autores evalúan sus resultados a través de la tarea de llenar los espacios en blanco con la traducción apropiada, que es esencialmente la evaluación del desempeño de la WSD.

En [2] se comparan cuatro métodos supervisados de WSD. El etiquetado del corpus de entrenamiento con los sentidos se lleva a cabo con los corpus paralelos alineados por palabras, del mismo modo como lo hacen Ng *et al*. [4].

Vickery *et al*. evalúan los resultados de la WSD aislada comparándolos con la línea base del sentido más frecuente. Dado que el objetivo final es medir la mejora de un sistema de TA, proponen una metodología de la evaluación especializada.

En [2] se afirma que los descodificadores disponibles, es decir los sistemas que encuentran la traducción más probable con respecto a algún modelo de probabilidad, actualmente no proporcionan una forma natural de incorporar los resultados de un sistema de la traducción de una palabra. Por lo tanto, para evaluar los resultados se utiliza la tarea de traducción simplificada. La traducción del texto se reduce a la tarea de "llenado de los espacios en blanco": en un texto paralelo se reemplazan las traducciones de las palabras ambiguas con los espacios en blanco. De esta manera la tarea de evaluación de la TA se convierte en la evaluación del desempeño de la WSD.

Resumiendo, a pesar de una formulación diferente de la tarea de Vickery *et al*., sus métodos de la evaluación son muy similares a los presentados en las secciones 2 y 3.

Specia [8] introduce un nuevo enfoque híbrido de WSD realizado con las técnicas de la programación de la lógica inductiva. El módulo de WSD o, mejor dicho, de la desambiguación de la traducción, introducido tiene el objetivo de aumentar el desempeño de TA, ya que el resultado de este módulo de WSD es una variante de la traducción correcta de una palabra ambigua para un contexto dado.. El módulo utiliza las evidencias de los dos idiomas de un par de la traducción y se ha diseñado específicamente para el par portugués-inglés.

Para evaluar el desempeño del módulo, Specia compara el desempeño de su método de WSD contra el desempeño de la WSD de dos algoritmos de aprendizaje – los árboles de decisión y las máquinas de soporte vectorial (en inglés: *Support Vector Machines*, SVM) – en un conjunto de datos estándar.

A pesar de que según Specia este enfoque tiene como objetivo mejorar la TA, ninguna introducción real de su método de WSD en un sistema de TA se llevó a cabo. Por lo tanto, ningunos métodos para la evaluación desarrollados específicamente para la WSD-para-la-TA no se han presentado.

Carpuat y Wu realizaron una serie de trabajos [9, 27–28] en la integración de los resultados de la WSD en los sistemas de SMT. En [9] y [28] efectivamente se pone en práctica la incorporación de un módulo de WSD en un sistema de SMT.

En su trabajo de 2005 [28] Carpuat y Wu muestran que los modelos de SMT actuales tienen limitaciones en la WSD, en comparación con los modelos dedicados a

la WSD, y que la SMT debería beneficiarse de las predicciones hechas por los módulos de WSD.

Para aclarar, en [28], no se introduce un método de WSD novedoso ni se mejoran los existentes. En cambio, se analizan las características de los modelos de SMT relevantes para la tarea de WSD y se experimentan con aplicaciones de los sistemas de SMT a la tarea de WSD.

La evaluación del desempeño se realiza en un conjunto de datos estándar de la tarea de muestra léxica china del Senseval-3. Los modelos de WSD supervisados se entrenaron con el inventario de sentidos HowNet. Con el fin de evaluar las predicciones del modelo de SMT al igual que cualquier modelo de WSD, los autores tuvieron que mapear las traducciones inglesas sobre las acepciones de HowNet.

Resumiendo, los autores lograron reducir sus tareas de la evaluación a la tarea de la comparación de diferentes sistemas de WSD en un conjunto de datos estándar con un inventario de sentidos mutuo.

Su otro trabajo de 2005 [27] ya presenta una integración real de los resultados de la WSD en un sistema de SMT. Este trabajo proporciona un resultado decepcionante: el uso de un modelo de WSD del estado-del- arte en chino para elegir los candidatos de la traducción para un sistema típico de SMT de IBM *no* mejora la calidad de la traducción considerablemente comparando con la del sistema de SMT por sí solo.

En [27] se presentan dos enfoques a la integración de los resultados de la WSD en un sistema de SMT. En el primer enfoque, las predicciones del sentido hechas por el algoritmo de WSD se usan para restringir el conjunto de los posibles sentidos en inglés considerados por el decodificador para cada una de las palabras elegidas para el experimento. En el segundo, las predicciones de la WSD se usan para el pos-procesamiento de la salida del sistema de SMT: en cada frase de salida, la traducción de la palabra observada es reemplazada directamente por la predicción de la WSD. Cuando el sistema de WSD predice más de un candidato, solo una traducción se elige al azar. En ambos casos se utiliza el mismo sistema de WSD basado en el modelo que logró el mejor desempeño en la tarea de la muestra léxica china del Senseval-3, Kernel PCA. El sistema de SMT también no se cambia durante todo el experimento.

Dado que la tarea de Carpuat y Wu en [27] no fue mejorar algún modelo de WSD sino investigar la influencia de la WSD en la calidad de la SMT, los autores emplean un método de evaluación utilizado para evaluar la traducción automática: el sistema de la evaluación de la traducción automática BLEU. Los autores comparan los resultados de las traducciones realizadas por el elegido sistema de SMT por sí solo y por el mismo sistema de SMT con un módulo de WSD integrado.

El último trabajo de la serie de los trabajos de Carpuat y Wu en el área de la integración de los resultados de la WSD en un sistema de SMT que revisamos es el trabajo [9] de 2007. En [9] se muestra que WSD sí mejora la calidad de la traducción de un modelo de SMT típico basado en frases. Sin embargo, para lograr esta mejora, los autores tuvieron que redefinir la tarea de WSD para que coincida exactamente con la misma tarea con la cual se enfrentan los sistemas de SMT basados en frases —a saber, la desambiguación de la traducción de las frases.

Por lo tanto, en este trabajo los autores efectivamente presentan un modelo de WSD novedoso, que ellos llaman “el modelo de la desambiguación del sentido de las frases completas compuestas por varias palabras”. En este modelo, en lugar de considerar una sola palabra como un objeto de la WSD, se desambiguan las frases

enteras con un vocabulario de la SMT como el inventario de sentidos. En este caso el modelo de WSD utiliza las mismas definiciones de sentidos y los datos de entrenamiento que el modelo de SMT.

La evaluación se lleva a cabo con dos estándares tareas de traducción de chino a inglés. Dado que el objetivo de Carpuat y Wu no es evaluar el desempeño de su modelo de WSD modificado, sino la calidad de la traducción, ellos utilizan la metodología estándar de evaluación de la TA y no evalúan la exactitud del modelo de WSD por sí sola. A diferencia de [27], donde se usa sólo una métrica estándar de evaluación de traducción automática BLEU, en [9] los resultados de la SMT se evalúan con ocho métricas diferentes.

Chan y Ng [10], al igual que Carpuat y Wu [27], incorporan un sistema de WSD del estado-del-arte en un sistema de SMT del estado-del-arte, que se llama Hiero. Pero a diferencia de [27] Chan y Ng logran obtener la mejora considerable en el desempeño del sistema de SMT en una tarea real de traducción.

Ya que su trabajo también tiene como objetivo mejorar la calidad de la TA en lugar de aumentar la precisión de la WSD, para evaluar sus resultados Chan y Ng usan la métrica de evaluación de la TA más popular: BLEU.

Resumimos, que la evaluación de los resultados de los trabajos en la rama de WSD para TA puede realizarse en dos modos:

- evaluar el desempeño del modelo de WSD directamente usando los métodos descritos en las secciones 2 y 3,
- evaluar la calidad final de la traducción de los sistemas de TA modificados con los módulos de WSD por medio de las métricas de evaluación de la TA estándares como BLEU.

5 Discusión y las conclusiones

La amplia comparación de una variedad de los trabajos en la mejora del desempeño de la WSD hecha en el artículo presente nos permite obtener varias conclusiones.

Basándonos en la revisión presentada en el artículo, podemos afirmar que las formas de la evaluación de la mejora de un modelo de WSD dependen de las modificaciones que llevan la dicha mejora y de objetivos del uso de la WSD. Los tipos de modificaciones de loa modelos de WSD pueden ser:

- la modificación del mismo algoritmo de WSD o introducción de un modelo o algoritmo de WSD novedoso como en los trabajos [2, 3, 8, 27],
- la modificación del inventario de sentidos contra el que se realiza la desambiguación como en los trabajos [1, 4, 15, 18, 19],
- varias modificaciones en el modelo de WSD con el fin de mejorar el desempeño de algún sistema de PLN más general, por ejemplo, los sistemas de TA, como en los trabajos [2, 9, 10, 27].

Se debe notar que no todos los trabajos, que clasifican sus modificaciones de una cierta manera, de hecho realizan la modificación declarada. Como podemos ver del análisis de la investigación conducida en el trabajo [8], la autora declara su modelo de WSD como uno diseñado especialmente para la TA, pero finalmente no propone

método para integrarlo en un sistema de SMT ni metodología de la evaluación de la mejora de la traducción debida a su modelo de WSD.

Diferentes tipos de modificaciones de la WSD requieren diferentes metodologías de la evaluación.

1. Así los modelos de WSD que modifican el mismo algoritmo de WSD o introducen un modelo o algoritmo de WSD novedoso para evaluarse deben comparar su desempeño contra los mejores métodos de WSD en la misma tarea (un conjunto de datos estándar) con el mismo inventario de sentidos para todos los métodos.
2. Los modelos de WSD que modifican el inventario de sentidos deben comparar el desempeño de por lo menos un algoritmo de WSD existente usando el inventario original y modificado en un conjunto de datos estándar.
3. Los modelos de WSD que tienen como su objetivo la mejora de un sistema de PLN más general deben emplear los métodos para medir el desempeño específicamente de ese sistema.

En todos los puntos también es preferible incluir como una posición de la comparación una modificación aleatoria con las mismas características como lo hacen Snow *et al*. [1].

Agradecimientos. Este trabajo fue realizado con el apoyo parcial del Gobierno de México (SNI, COFAA-IPN, PIFI-IPN, SIP-IPN 20113295 y 20111146, CONACYT 50206-H), CONACYT-DST India (proyecto "*Answer Validation through Textual Entailment*"), Gobierno del DF, México (ICYT PICCO10-120), Proyecto Europeo WIQ-EI 269180.

6 References

1. Snow, R., Prakash, S., Jurafsky, D., y Ng, A. Y.: Learning to Merge Word Senses. EMNLP-CoNLL 2007, pp. 1005–1014. ACL (2007)
2. Vickrey, D., Biewald, L., Teyssier, M. y Koller, D.: Word-sense disambiguation for machine translation. EMNLP 2005, pp. 771–778. ACL (2005)
3. Diab, M.: Word sense disambiguation within a multilingual framework. Ph.D. dissertation. University of Maryland, College Park, College of Park, MD (2003)
4. Ng, H. T., Wang, B., y Chan, Y. S.: Exploiting parallel texts for word sense disambiguation: an empirical study. ACL 2003, vol. 1, pp. 455–462. ACL (2003)
5. Ide, N.: Cross-lingual sense determination: Can it work? En: Computers and the Humanities, vol. 34, 1–2, pp. 223–234. Springer (2000)
6. Ide, N.: Making senses: Bootstrapping sense-tagged lists of semantically-related words. En: Gelbukh, A. (Ed.). Computational Linguistics and Intelligent Text Processing. LNCS, vol. 3878, pp. 13–27. Springer (2006)
7. Ide, N. y Wilks, Y.: Making sense about sense. En: Agirre, E. y Edmonds, P. (Eds.) Word Sense Disambiguation: Algorithms and Applications, pp. 47–73. Springer (2006)
8. Specia, L.: A hybrid relational approach for WSD—first results. COLING/ACL 2006 Student Research Workshop, pp. 55–60. ACL (2006)
9. Carpuat, M. y Wu, D.: Improving statistical machine translation using word sense disambiguation. EMNLP-CoNLL 2007, pp. 61–72. ACL (2007)

10. Chan, Y. S. y Ng, H. T.: Word sense disambiguation improves statistical machine translation. ACL 2007, pp. 33–40. ACL (2007)
11. Stark, M. M. y Richard F. Riesenfeld, R.F.: WordNet. An electronic lexical database. En: Memorias de 11th Eurographics Workshop on Rendering. MIT Press (1998)
12. Agirre, E. y Edmonds, P.: Word Sense Disambiguation: Algorithms and Applications. Springer (2006)
13. Gale, W.A., Church, K., y Yarowsky, D.: Estimating upper and lower bounds on the performance of word-sense disambiguation programs. En: Memorias de the 30th Annual Meeting of the Association for Computational Linguistics, pp. 249–256. ACL (1992)
14. Kilgariff, A.: Word senses. En: Agirre, E. y Edmonds, P. (Eds.) Word Sense Disambiguation: Algorithms and Applications, pp. 29–46. Springer (2006)
15. Navigli, R.: Meaningful clustering of senses helps boost word sense disambiguation performance. En: Memorias de the 44th Annual Meeting of the Association for Computational Linguistics joint with the 21st International Conference on Computational Linguistics, pp. 105–112. ACL (2006)
16. Navigli, R.: Word Sense Disambiguation: a Survey. En: ACM Computing Surveys, vol. 41 (2), pp. 1–69. ACM Press (2009)
17. Weaver, W.: Translation. En: Locke, W. N. y Booth, A.D. (Eds.) Machine Translation of Languages: Fourteen Essays. MIT Press, Cambridge (1949)
18. Ide, N., Erjavec, T. y Tufis, D.: Automatic sense tagging using parallel corpora. En: Memorias de 6th Natural Language Processing Pacific Rim Symposium, pp. 212–219 (2001)
19. Ide, N., Erjavec, T. y Tufis, D.: Sense discrimination with parallel corpora. En: Memorias de ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, pp. 54–60. ACL (2002)
20. Snyder, B. y Palmer, M.: The English all-words task. En: Memorias de ACL 2004 Senseval-3 Workshop, pp. 41–43. ACL (2004)
21. Decadt, B., Hoste, V., Daelemans, W., Bosch, A.: Gambl, genetic algorithm optimization of memory-based WSD. En: Memorias de ACL/SIGLEX Senseval-3, pp. 108–112. (2004)
22. Mihalcea, R. y Faruque, E.: Senselearner: Minimally supervised word sense disambiguation for all words in open text. En: Memorias de ACL/SIGLEX Senseval-3, 155–158. (2004)
23. Yuret, D.: Some experiments with a naïve bayes wsd system. En: Memorias de ACL/SIGLEX Senseval-3, pp. 265–268. ACL (2004)
24. Strapparava, C., Gliozzo, A., Giuliano, C.: Pattern abstraction and term similarity for word sense disambiguation, 229–234. En: Memorias de ACL/SIGLEX Senseval-3. ACL (2004)
25. Navigli, R. y Velardi, P.: Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. En: IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(7), pp. 1075–1086. IEEE (2005)
26. Lee, Y. K. y Ng, H. T.: An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. En: Memorias de the 2002 Conference on Empirical Methods in Natural Language Processing, pp. 41–48. ACL (2002)
27. Carpuat, M. y Dekai Wu, D.: Word sense disambiguation vs. statistical machine translation. En: Memorias de the annual meeting of the association for computational linguistics, pp. 387–394. ACL (2005)
28. Carpuat, M. y Dekai Wu, D.: Evaluating the word sense disambiguation performance of statistical machine translation. En: Memorias de the Second International Joint Conference on Natural Language Processing (IJCNLP), pp. 122–127. IJCNLP (2005)
29. Ledeneva, Y. y Sidorov, G. Recent Advances in Computational Linguistics. Informatica. International Journal of Computing and Informatics, 34 (2010) pp. 3–18 (2010)