Resolución Automática de la Ambigüedad Léxica fundamentada en el Modelo del Espacio Vectorial usando Ventana Contextual Variable

L. ALFONSO UREÑA LÓPEZ MANUEL GARCÍA VEGA
Departamento de Informática.
Universidad de Jaén
Avda. Madrid 35, 23071 Jaén. Spain
e-mail: {laurena,mgarcia}@ujaen.es

MANUEL DE BUENAGA RODRÍGUEZ JOSÉ MARÍA GÓMEZ HIDALGO
Departamento de Inteligencia Artificial
Universidad Europea de Madrid
28670 - Villaviciosa de Odón. Madrid. Spain
e-mail: {buenaga,jmgomez}@dinar.esi.uem.es

Resumen
La resolución automática de la ambigüedad léxica de términos polisémicos es una tarea útil y compleja para muchas aplicaciones del procesamiento del lenguaje natural. Presentamos un nuevo enfoque basado en la utilización del modelo del espacio vectorial y una colección de entrenamiento, de amplia cobertura, como recurso lingüístico. Este enfoque utiliza el concepto de ventana contextual, un conjunto variable de términos como contexto local. Hemos probado que nuestro programa desambiguador del sentido de las palabras, sobre un gran conjunto de documentos, consigue una alta precisión, con un tamaño variable de ventana contextual, en la resolución de la ambigüedad léxica.
Palabras Clave*: Desambiguación del sentido de las palabras (WSD), Ventana Contextual, WordNet, SemCor, Corpus de Texto, Bases de Datos Léxicas.*

Abstract
The resolution of lexical ambiguity of polysemics words is a complex and useful task for many natural language processing applications. We present a new approach for word sense disambiguation based in the vector space model and a widely available training collection as linguistic resource. This approach uses a contextual windows (variable set of terms like local context). We have tested our disambiguator algorithm on a large documents collection, achieving high precision in the resolution of lexical ambiguity.
Key Words*: Word Sense Disambiguation (WSD), Contextual Windows, WordNet, SemCor, Copora, Lexical Database.*

1 INTRODUCCIÓN

Un problema importante en el procesamiento del lenguaje natural es determinar el sentido o significado de palabras ambiguas en un determinado contexto. Esta tarea de desambiguación es compleja para la mayoría de las aplicaciones del procesamiento del lenguaje natural. El perfeccionamiento en la identificación del sentido correcto de una palabra puede

mejorar el comportamiento de sistemas de traducción automática (Wilks 1990), sistemas de recuperación de información (Sanderson 1996); o utilizarse en tareas especificas como la restauración de acentos de palabras en el procesamiento de textos (Yarowsky 1994).

El problema de la desambiguación no es nuevo: Gale, Church y Yarowsky (Gale 1992) citan trabajos que se remontan a los años 50, si bien su posible utilización en aplicaciones sobre textos de dimensiones reales, apenas se plantea. Sin embargo, en 1986, Lesk (Lesk 1986) construye un desambiguador con técnicas similares a las utilizadas en los sistemas de recuperación de información, con mayores posibilidades de escalabilidad. Desde el trabajo de Lesk hasta la actualidad, se han ideado enfoques muy diversos para la desambiguación automática (Yarowsky 1992-1994; Miller 1994; Bruce 1994; Hwee 1996; Rigau 1997; Ureña 1997).

Realizaremos una clasificación genérica dependiendo del *recurso léxico* utilizado: diccionarios, o córpora de entrenamiento. Los sistemas basados en diccionarios utilizan la información que aparece en las definiciones de las distintas acepciones del término a desambiguar (Lesk 1986), y otros recursos léxicos como thesaurus (p. e. Roget's Thesaurus) (Yarowsky 1992) o bases de datos léxicas (p. e. WordNet) (Agirre 1996). En los enfoques basados en córpora de entrenamiento, se utilizan córpora etiquetados (p. e. SemCor) (Bruce 1994; Yarowsky 1996). En estos enfoques las técnicas de análisis varían según los casos, empleando técnicas bayesianas, probabilísticas e incluso redes neuronales. En nuestro enfoque nos hemos basado en la utilización de córpora de entrenamiento, ya que, con este tipo de información, existe una cierta evidencia experimental que proporciona mejores resultados (Yoshiki 1994). En concreto, utilizamos SemCor (Semantic Concordance), debido a su disponibilidad y amplia cobertura. SemCor está compuesto por el Brown Corpus, etiquetado manualmente con los sentidos de las palabras definidas en WordNet (Miller 1990, 95).

Algunos de los problemas que se presentan en los enfoques basados en córpora de entrenamiento son: primero, la mayoría de ellos incluyen un gran número de decisiones "ad-hoc", segundo, la escalabilidad de los métodos no queda clara en algunos casos, y tercero, la evaluación de su comportamiento se realiza para un número reducido de términos.

En este trabajo presentamos un nuevo enfoque para desambiguar el sentido de las palabras (WSD)[1] utilizando un corpus de entrenamiento basado en el Modelo del Espacio Vectorial (MEV) y estructurado en ventanas contextuales, entendidas como un conjunto variable de términos que es usado como contexto local del término a desambiguar. Este modelo (MEV), ampliamente fundamentado, tanto teórica como experimentalmente en el campo de la Recuperación de Información (Salton 1983, 89; Frakes 1992), permite el desarrollo de sistemas prácticos y eficientes. Presentamos además, una evaluación de nuestro método de desambiguación, para un conjunto extenso de documentos y de términos diferentes del Brown Corpus, con los sentidos afinados de WordNet, que ha determinado que el tamaño óptimo de ventana contextual es variable y se ajusta al tamaño del párrafo donde está incluido el término. Además, nuestro enfoque puede facilitar en el futuro, la integración de ambos recursos léxicos (diccionarios y córpora de entrenamiento), como ya hemos realizado en categorización de textos (Gómez 1997).

Este trabajo está organizado como sigue. Primero, introducimos la tarea para la resolución de la ambigüedad léxica y los recursos que utilizamos. Seguidamente, describimos el modelo en el que integramos estos elementos. Después de esto, estudiamos el proceso de entrenamiento, seguido de la descripción del proceso WSD. A continuación, presentamos nuestra evaluación estudiando e interpretando los resultados, y finalmente, describimos nuestras conclusiones y líneas futuras.

[1] Denominado también resolución de la ambigüedad léxica, discriminación del sentido de las palabras, selección del sentido de las palabras o identificación del sentido de las palabras.

2 DESCRIPCIÓN DE LA TAREA

La entrada de nuestro programa WSD consta de un conjunto de documentos en lenguaje natural. En la salida, los términos que componen dichos documentos de entrada, son etiquetados con el significado correcto, de acuerdo con los sentidos proporcionados por WordNet. El sistema hace uso de la información contenida en los documentos para determinar el significado. Como es lógico, no todos los términos tendrán el mismo número de acepciones, sino que éste será variable, dependiendo de la palabra en cuestión. Los sentidos están representados con etiquetas numéricas que codifican el sentido en WordNet.

El recurso más ampliamente usado para WSD, es la colección de entrenamiento. Una *colección de entrenamiento* es un conjunto de documentos etiquetados manualmente, con cada palabra acompañada del significado con que es utilizada. La colección de entrenamiento permite al sistema deducir en nuevos documentos los significados de las palabras. En nuestro trabajo hemos utilizado SemCor, dada su libre disposición y su utilización en trabajos relacionados, lo que nos ha facilitado la interpretación del comportamiento de nuestro sistema.

3 MODELO DEL ESPACIO VECTORIAL PARA LA RESOLUCIÓN DE LA AMBIGÜEDAD LÉXICA

Nuestro modelo de desambiguación se basa en el modelo del espacio vectorial (Salton 1983). El modelo del espacio vectorial ha sido utilizado en muchos trabajos de recuperación de información (Lewis 1992; Salton 1983, 89), así como en otros de categorización de textos (Gómez 1997). Basándonos en estas experiencias, presentamos una adaptación del modelo del espacio vectorial a la desambiguación de textos y los caminos seguidos para calcular algunos elementos del modelo.

Nosotros utilizamos el MEV para representar el lenguaje natural por medio de vectores de pesos. Cada peso representa la importancia de un término, en relación con un determinado sentido en la expresión del lenguaje natural. Cada término s_{ji} queda representado o indexado por un vector de dimensión *m,* con los pesos asignados a cada uno de los términos de indexación. El término *i* con sentido *j*, queda representado con el peso del término, así como con los pesos de los términos circundantes.

$s_{ji}=\langle ws_{j1}, ws_{k1}, ..., ws_{kn}\rangle$ ws_{kc} peso de la palabra circundante c al término s_{ji}

Para el procesamiento de los textos a desambiguar, se procede análogamente: la representación de una consulta de un término c_k, se realiza mediante un vector de pesos asociados a los términos.

$c_k=\langle wc_1, wc_{k1}, ..., wc_{kn}\rangle$ wc_{kc} peso de la palabra circundante c al término c_k

El MEV supone que la similitud semántica entre los objetos representados viene dada por el coseno del ángulo que forman sus vectores. Para desambiguar un término c_k, calculamos la similitud entre el contexto en el que aparece y las definiciones de los términos. Se selecciona el término que presenta mayor similitud con el contexto, con arreglo a la fórmula:

$$sim(s_{ji}, c_i) = \frac{\sum_{i=1}^{m} ws_{ji} \cdot wc_i}{\sqrt{\sum_{i=1}^{m} ws_{ji}^{\,2} \cdot \sum_{i=1}^{m} wc_i^{\,2}}}$$

4 ENTRENAMIENTO CON VENTANA CONTEXTUAL VARIABLE

Representamos cada término del córpora de entrenamiento, por medio de un vector, cuyas componentes son: el peso del término en el párrafo; y los pesos de los términos que constituyen lo que hemos denominado *ventana contextual*. Con este concepto, hacemos referencia a las palabras que circundan al término a desambiguar, es decir, a las palabras que están en su contexto, ya que pueden suministrar información acerca del sentido utilizado. Así, para cada uno de los nombres de la colección de entrenamiento, calcularemos su ventana contextual. El programa desplaza la ventana desde el principio hasta el final de todos los documentos que contiene el córpora de entrenamiento, considerando en cada desplazamiento un nombre, y como palabras de contexto, cada una de las palabras circundantes. De esta manera, se construyen tantas ventanas contextuales como términos lexicográficamente distintos y con diferentes sentidos existan en la colección. El tamaño de la ventana contextual, será variable, como demostraremos, adaptándose al tamaño del párrafo donde aparece el nombre.

Una vez realizado esto, se construyen los vectores para la colección de entrenamiento y se calculan los pesos para los distintos términos de manera análoga a Salton (Salton 1983):

$$ws_{ji} = t_{ji} * w_i \qquad w_i = \log_2(n/f_i)$$

Donde t_{ji} es la frecuencia del término j con sentido i en la ventana contextual, *n* es el número de sentidos de término i y f_i es el número de ventanas contextuales donde aparece el término *i*. La dimensión del espacio vectorial es variable y está en función del número de sentidos que tenga la palabra a desambiguar.

5 DESAMBIGUACIÓN DEL SENTIDO DE LAS PALABRAS

La técnica utilizada para la resolución automática de la ambigüedad léxica, se enmarca dentro de la adquisición, con representaciones contextuales, del significado de las palabras. Una representación contextual es una caracterización del contexto lingüístico en el que una palabra expresa un determinado sentido (Miller91). Por tanto, nuestro algoritmo trata de encontrar y representar características contextuales, a través de ventanas contextuales de tamaño variable. En nuestros experimentos hemos considerado el párrafo, tal y como está definido en SemCor, como ventana contextual máxima, entendiéndolo pues como unidad semántica.

Con la consulta (que constituye el córpora de prueba a desambiguar) actuamos análogamente. Primero, calculamos las ventanas contextuales de la colección de prueba, que nos dará el contexto válido de los términos a desambiguar y, construyendo después los vectores, calculamos la similitud. Ésta nos proporcionará el sentido correcto en relación con el término a desambiguar de la colección de prueba.

6 EVALUACIÓN

La evaluación de sistemas WSD ha supuesto un problema en la investigación de la desambiguación, de hecho, algunos desambiguadores han sido evaluados sólo a través de pruebas manuales sobre un grupo reducido de palabras. La evaluación es muy heterogénea. En numerosos trabajos se han utilizado diferentes métricas y colecciones de prueba. Para nuestro trabajo hemos adoptado las basadas en el campo de la evaluación de los sistemas de recuperación de información (Salton 1983; Frakes 1992), así como una colección de prueba de amplia cobertura y libre disposición. Para evaluar el rendimiento, hemos realizado un test sobre un conjunto de documentos, que será comparado con las pruebas realizadas sobre un algoritmo *línea base*. Éste asigna al término a desambiguar el sentido más frecuente en el corpus (Boguraev 1996).

6.1 Métricas de Evaluación

Hemos utilizado la precisión como métrica básica para computar la efectividad de nuestros experimentos. El cálculo puede ser realizado utilizando *macroaveraging* y *microaveraging* (Lewis 1992). La Precisión puede ser definida como el cociente entre el número de términos desambiguados satisfactoriamente y el número de términos desambiguados. El macroaveraging consiste en calcular la precisión para cada uno de los términos, y luego calcular la media para cada uno de ellos; y el microaveraging en calcular un solo valor de precisión medio para todos los términos.

$$P_{macroavg} = \frac{\sum_{i=1}^{n} P_i}{n} \quad P_i = \text{Precisión del término } i \qquad P_i = \frac{dc_i}{dc_i + di_i}$$

Donde dc_i es el número de desambiguaciones correctas del término i, di_i el número de desambiguaciones incorrectas del término i y n el número de términos desambiguados.

$$P_{microavg} = \frac{tdc}{tdc + tdi}$$

Siendo *tdc* el número de términos desambiguados correctamente y *tdi* el número de términos desambiguados incorrectamente.

6.2 Colección de Prueba

Para nuestros experimentos, hemos utilizado SemCor (Miller 1993), que además de un corpus de texto, es un lexicón, donde cada palabra en el texto hace referencia a su correcto significado en él. Puede definirse bien como un corpus, en el que las palabras han sido etiquetadas sintáctica y semánticamente, o como un lexicón, en el cual las frases de ejemplo pueden ser encontradas por varias definiciones. SemCor es el Brown Corpus donde sólo los nombres, verbos, adjetivos y adverbios son etiquetados semánticamente con los sentidos de WordNet. Las palabras etiquetadas no tienen punteros a WordNet. Las palabras (tales como preposiciones, determinantes, pronombres, verbos auxiliares, etc.) y caracteres no alfanuméricos, interjecciones y términos coloquiales no son etiquetados.

El algoritmo produce un fichero de resultados, para cada uno de los documentos seleccionados aleatoriamente, con los sentidos inferidos para que puedan ser comparados automáticamente con los ficheros originales.

6.3 Tamaño de la ventana contextual

Uno de los objetivos de nuestros experimentos fue decidir entre diferentes unidades contextuales: frase, párrafo o bien otros tamaños de ventanas.

Para ello, hemos seleccionado aleatoriamente cuatro documentos o textos de SemCor considerados individualmente: br-a14, br-j09, br-k11 y br-k14. Estos textos han representado el papel de ficheros de entrada (sin etiquetas).

Para comparar nuestro algoritmo, como hemos comentado anteriormente, hemos decidido implementar un algoritmo línea base (Boguraev 1995), y confrontarlo con nuestros experimentos. Primeramente, hemos tomado como unidad contextual la frase, tal y como la define SemCor.

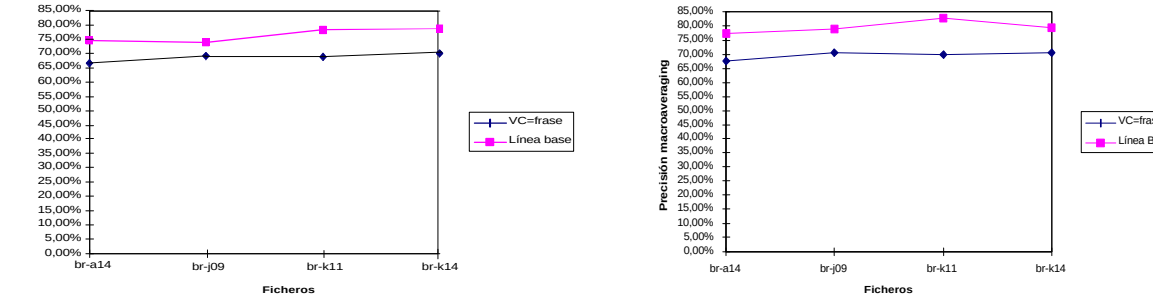

Figura 1: Microaveraging y macroaveraging para una ventana contextual de tamaño frase.

La precisión *microaveraging* y *macroaveraging* obtenida por nuestro algoritmo, utilizando como tamaño de ventana contextual la frase, es más baja que la obtenida por el algoritmo línea base (ver Figura 1). La unidad contextual frase no tiene un tamaño fijo, sino variable en SemCor, puesto que cada frase puede tener un número distinto de términos. De este experimento se puede concluir que la unidad contextual frase no es el tamaño óptimo.

En segundo lugar hemos tomado diferentes tamaños de ventana contextual como adquisición de conocimiento (de 10 a 60 términos) y hemos realizado los experimentos para los mismos ficheros seleccionados anteriormente. Se muestra en la Figura 2 la precisión *microaveraging* y *macroaveraging*.

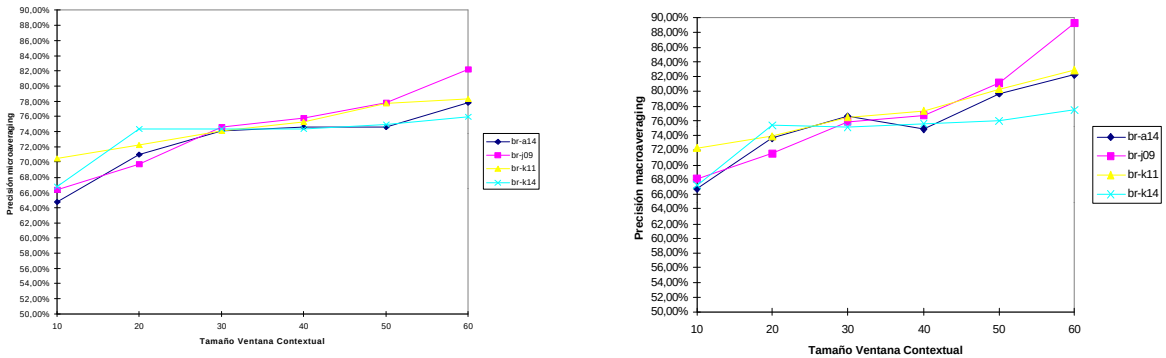

Figura 2: Microaveraging y macroaveraging para una ventana contextual con un tamaño en el intervalo (10,60).

Podemos observar que, conforme aumenta el tamaño de la ventana contextual, mayor es la precisión de nuestro algoritmo. Hemos considerado como tamaño máximo 60 términos circundantes, debido a que éste es el número significativo máximo computado en SemCor por nuestro algoritmo y a partir de este tamaño se obtienen resultados idénticos.

Se obtiene empíricamente que cuanto mayor es la ventana contextual más precisión alcanza nuestro algoritmo, puesto que se adquiere mejor contexto, así cuando el tamaño de la ventana contextual se aproxima al párrafo se consigue la mayor precisión.

Algoritmos WSD	*Macroaveraging*	*Microaveraging*
	Precisión	
Basado MEV	83%	78,6%
Línea Base	81%	76%

Tabla 1. Resultados totales de nuestros experimentos.

6.4 Resultados e Interpretación

La precisión obtenida por nuestro algoritmo es alta. En la Tabla 1 podemos observar las medias macroaveraging y microaveraging para la precisión, obtenidas en nuestros experimentos. De nuevo hemos utilizado los mismos ficheros que en el cálculo del tamaño óptimo de la ventana contextual: br-a14, br-j09, br-k11 y br-k14

Destacamos que hemos probado nuestro algoritmo con **todos** los nombres en un subconjunto del córpora no restringido, haciendo distinción entre el gran número de sentidos de WordNet.

7 Conclusiones y Futuros Trabajos

En este trabajo hemos presentado un nuevo enfoque para desambiguar el sentido de las palabras de SemCor, combinando el modelo del espacio vectorial con un algoritmo basado en el aprendizaje. Hemos utilizado el Modelo del Espacio Vectorial, ampliamente usado en recuperación de información, y el concepto de ventana contextual de tamaño variable. Hemos obtenido buenos resultados en la resolución de la ambigüedad de términos polisémicos, como son los términos de SemCor etiquetados con los sentidos de WordNet.

El algoritmo está teóricamente fundamentado y ofrece una medida general de la relatividad semántica para algún número de nombres en un texto. En el experimento, el algoritmo desambiguó cuatro textos de SemCor. Los resultados fueron obtenidos automáticamente comparando las etiquetas de SemCor con los calculados por el algoritmo, lo cual permite la comparación con otros métodos de desambiguación. La evaluación del comportamiento del algoritmo proporciona resultados positivos, a pesar de la dificultad de la tarea y de la gran variedad de sentidos definidos en WordNet por palabra.

Por otro lado, nuestro enfoque puede facilitar en el futuro, la integración de varios recursos léxicos, como diccionarios y córpora de entrenamiento. Creemos que combinando las definiciones de los términos del diccionario con las ventanas contextuales construidas a partir de córpora de entrenamiento, obtendremos mejores resultados.

Referencias

Agirre E., Rigau G. 1996. "Word sense disambiguation using conceptual density". *In Proceedings of COLING'96.*

Boguraev B., Pustejovsky J. 1996. "Corpus processing for lexical acquisition". *A Bradford Book, Language, Speech and Communication Series,* The MIT Press.

Bruce R., Wiebe J. 1994. "Word sense disambiguation using decomposable models". *In Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics, (ACL' 94).*

Frakes, W., Baeza, R. 1992. *Information retrieval: data structures and algorithms,* Prentice Hall, London.

Gale W., Church KW., Yarowsky D. 1992. "Estimating upper and loewr bounds on the performance or word-sense disambiguation programs". In Proceedings of the ACL'92.

Gómez J.M., Buenaga De M. 1997. "Integrating a lexical database a training collection for text categorization". *In Proceedings of ACL'97.*

Hwee Tou Ng, Hian Beng Lee 1996. "Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based". *In Proceedings of ACL'96.*

Lesk M. 1996. Automatic sense disambiguation: how to tell a pine cone from an ice cream cone. *In Proceedings of the SIGDOC Conference.*

Lewis, D. 1992. "Representation and learning in information retrieval". *Ph.D. Thesis, Department of Computer and Information Science,* University of Massachusetts.

Miller G. 1990. "WordNet: An on-line lexical database". *An International Journal of Lexicography.*

Miller G., Charles G. 1991. "Contextual correlates of semantic similarity". *Language and Cognitive Processes.*

Miller G. Leacock C., Randee T. and Bunker R. 1993. "A Semantic concordance". *In Proceedings of the 3rd DARPA Workshop on Human Language Technology,* New Jersey.

Miller G. Chodorow M., Landes S., Leacock C. and Thomas R. 1994. "Using a semantic concordance for sense identification". *In Proceedings of the ARPA Human Language Technology.*

Rigau G., Atserias J., Agirre E. 1997. "Combining unsupervised lexical knowledge methods for word sense disambiguation". *In Proceedings of ACL' 97.*

Salton G., McGill, M.J. 1983. *Introduction to modern information retrieval*. McGraw-Hill.

Salton, G. 1989. *Automatic Text Processing: the transformation, analysis and retrieval of information by computer*. Addison Wesley.

Sanderson, M. 1996. "Word sense disambiguation and information retrieval". *Ph.D. Thesis, Department of Computing Science*, University of University of Glasgow.

Ureña López, L. A., García Vega, M., Buenaga Rodríguez, M., Gómez Hidalgo, J. M. 1997. "Resolución de la ambigüedad léxica mediante información contextual y el modelo del espacio vectorial". *Séptima Conferencia de la Asociación Española para la Inteligencia Artificial.* CAEPIA'97.

Wilks Y., Fass D., Guo C., McDonald J., Plate T., Slator B. 1993. "Providing machine tractable dictionary tools". *Machine Translation*.

Yarowsky D. 1992. "Word-sense disambiguation using statistical models of Roget's categories trained on large corpora". *In Proceedings of the 15th International Conference onr Computational Linguistics.*

Yarowsky D. 1994. "Decision lists for lexical ambiguity resolution: Aplication to accent restoration in Spanish and French". *In Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics*, (ACL' 94).

Yarowsky D. 1995. "Unsupervised word sense disambiguation rivaling supervised methods". *In Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics*, (ACL' 95).

Yoshiky N., Yoshihiko Nitta. 1994. "Co-ocurrence vectors from corpora vs. distance vectors from". *In Proceedings of COLING'94.*