Comparación léxica de corpus para generación de modelos de lenguaje

Luis Villaseñor[1], Manuel Montes[1], Manuel Pérez[1], Dominique Vaufreydaz[2]

[1]Laboratorio de Tecnologías de Lenguaje, Instituto Nacional de Astrofísica, Óptica y Electrónica,
Apdo. Post al 51 y 216, 72000, Puebla, Pue. México.
{villasen, mmontesg, mapco}@inaoep.mx

[2]Laboratorio CLIPS-IMAG, Universidad Joseph Fourier, Campus Scientifique,
BP 53, 38041 Grenoble cedex 9, Francia.
Dominique.Vaufreydaz@imag.fr

Resumen.

En este artículo se presenta un estudio para evaluar la riqueza léxica de un corpus específicamente recolectado para el entrenamiento de modelos de lenguaje estadísticos. Para ello se presenta un estudio comparativo entre un corpus oral –el corpus DIME– y un corpus recolectado de la Web para la construcción de modelos de lenguaje –el corpus WebDIME–. Los resultados de dicho análisis permiten identificar varios puntos débiles del corpus WebDIME. Básicamente, el coeficiente de diferencia es de 0.71, es decir, el porcentaje de ocurrencias de los términos en estos corpus difiere de manera importante, dado a que existen varios términos sobre o subrepresentados. Este conjunto de términos (las palabras críticas) representa cerca del 4% del total de palabras en el vocabulario.

Abstract.

In this paper, we present a study of the lexical richness of a corpus used for training statistical language models. We also present a comparative analysis of a given oral corpus – the DIME corpus – and a corpus obtained from the web – the WebDIME corpus. The results from this analysis are interesting because they identify several weaknesses of the WebDIME corpus. Basically, we obtained a difference coefficient of 0.71 between the corpora, indicating that the term occurrence proportions are considerably different in the two corpora. This result indicates that several terms of the DIME corpus are sub-represented or over-represented in the WebDIME corpus. The set of critical words (terms) corresponds to 4% of the complete vocabulary.

Palabras clave: modelo de lenguaje, análisis léxico, corpus.

1. Introducción

Uno de los principales componentes de un sistema de reconocimiento automático de voz es el modelo de lenguaje (ML). Este componente colabora en la selección del enunciado más apropiado dadas ciertas observaciones acústicas. Básicamente, el modelo de lenguaje captura el contexto de uso de un conjunto de palabras y con esta información dictamina la probabilidad de ocurrencia de una secuencia de dos o más términos. Existen básicamente dos tipos de modelos de lenguaje: los basados en gramáticas y los estadísticos. Los ML estadísticos son más flexibles y son capaces de capturar situaciones más

cercanas al lenguaje oral espontáneo donde no siempre las reglas del lenguaje escrito son respetadas. Esto es posible gracias a la capacidad de los ML estadísticos de explotar las propiedades estadísticas del lenguaje en contextos de dos, tres o más palabras.

Los ML estadísticos son calculados generalmente a partir de grandes corpus de texto delimitándolos por el tamaño del vocabulario, la longitud del contexto e incluyendo esquemas para tratamiento de palabras desconocidas [Jurafsky & Martin, 00]. Uno de los factores determinantes de un ML es el tamaño del corpus usado durante la fase de entrenamiento. Mientras más grande sea el corpus mayor será el número de contextos de uso de una palabra dada, y por ende, el ML obtenido será más robusto.

Ahora bien, los corpus recabados específicamente para el estudio de la comunicación hombre-máquina – utilizando métodos como el *Mago de Oz* –, son demasiado pequeños para entrenar ML pertinentes al contexto de aplicación. Una solución es utilizar la Web como fuente de datos [Gelbukh et al., 02, Vaufreydaz et al., 99]. Actualmente, mucha gente tiene acceso a Internet, ya sea desde su trabajo, escuela u hogar. Este mundo de personas no sólo consulta los documentos existentes sino que también participan activamente en la elaboración de nuevos contenidos enriqueciendo continuamente el conjunto de documentos disponibles. Dependiendo del contexto, los documentos presentan características muy diferentes. Una parte de estos documentos están elaborados usando un lenguaje familiar con un vocabulario simplificado que incluye expresiones cotidianas y, en ocasiones, hasta faltas gramaticales propias del lenguaje oral. Esto significa que usando los documentos disponibles en Internet podemos recolectar un corpus de grandes dimensiones, con una mezcla de texto correctamente escrito y texto libre más cercano al lenguaje oral. Por supuesto, no deja de ser texto escrito donde fenómenos lingüísticos del habla oral estarán mal representados.

Dado que el ML es un reflejo del corpus de entrenamiento ¿cómo podemos medir la semejanza léxica de un corpus recolectado de Internet con respecto a un corpus de referencia específico del dominio?. El presente artículo presenta un estudio comparativo a nivel léxico para responder esta pregunta. La siguiente sección describe las colecciones de datos usadas en este estudio (DIME y WebDIME). La tercera sección introduce algunos conceptos sobre el análisis léxico, explica el estudio comparativo y muestra los resultados obtenidos. La última sección, enlista las conclusiones y perspectivas de este estudio.

2. Las colecciones DIME y WebDIME

2.1 El corpus DIME

Bajo el contexto del proyecto DIME *"Diálogos Inteligentes Multimodales en Español"* [Pineda et al., 02] se realizó la recolección de un corpus multimodal [Villaseñor et al., 01]. Este corpus – también llamado DIME - consiste en un conjunto de diálogos recolectados a través del escenario del Mago de Oz [Bernsen et al., 98]. Un escenario del Mago de Oz consiste, generalmente, de una persona (el mago) quien toma el papel del sistema computacional y, un interlocutor (el sujeto) a quien se le solicita la resolución de una tarea en el dominio de interés. Todo esto bajo una plataforma computacional, acercándose lo más posible a una interacción hombre-máquina real. Las características del escenario pueden variar dependiendo del objetivo del estudio. En particular, el objetivo del corpus DIME fue la adquisición de diálogos multimodales. Para ello, se pensó en una aplicación sencilla, fácil de llevar a cabo por la mayoría de la gente, con un cierto grado de dificultad que implicara la presencia de un asistente, y donde los fenómenos multimodales fueran fácilmente observables: el diseño de cocinas. Un fragmento del tipo de diálogos recolectados puede observarse en la tabla 1.

```
utt37:  u:  me puedes poner e[l] <sil> el tercero
                junto a la estufa <sil> éste ?
utt38:  s:  éste junto a la estufa ?
utt39:  u:  sí / por favor
utt40:  s:  okey
utt41:  s:  <ruido> <no-vocal> así está bien ?
utt42:  u:  <ruido> <sil> sí / así está bien
```

Tabla 1. Fragmento de un diálogo del corpus DIME

En estos diálogos los participantes no tienen contacto visual entre ellos, únicamente contacto por voz y pueden referirse a los muebles de la cocina a través de una interfaz gráfica y un dispositivo de designación directa (i. e. el ratón). La tabla 2 resume los datos principales del corpus DIME.

Total (31 diálogos)		*En promedio (por diálogo)*	
27459	instancias de términos	886	instancias de términos
5779	elocuciones	185	elocuciones
3606	turnos	115	turnos
7:10	horas	14	minutos
1129	formas léxicas		

Tabla 2. Datos principales del corpus DIME

2.2 El corpus WebDIME

El corpus WebDIME se recolectó siguiendo las ideas de [Vaufreydaz et al., 99]. Utilizando el robot Web *CLIPS-Index* se recolectó una enorme cantidad de documentos en español de la Web. Esta colección es de cerca de 30 gigabytes de documentos HTML. Después de transformar los documentos HTML a una forma textual más apropiada, una serie de filtros son utilizados para homogeneizar las formas léxicas: transformación de mayúsculas a minúsculas, cifras a texto, etc. Finalmente, basados en el vocabulario de nuestra tarea, aplicamos la técnica de *bloques mínimos.* Un bloque mínimo de orden *n* es una secuencia de al menos *n* palabras consecutivas, todas ellas dentro del vocabulario dado. La tabla 3 muestra un ejemplo de esta técnica. En ella se muestran las secuencias de salida de la frase "pon el teclado debajo de la mesa" suponiendo que la palabra *teclado* no se encuentra en el vocabulario.

Longitud del bloque mínimo	*Secuencias de salida*
2	pon el debajo de la mesa
3	debajo de la mesa
4	debajo de la mesa
5	Ø

Tabla 3. Secuencias de salida de la frase "pon el teclado debajo de la mesa"

Como resultado de aplicar la técnica de bloques mínimos al corpus general obtuvimos un corpus específico para el diseño de cocinas: el corpus WebDIME. La tabla 4 muestra un fragmento del tipo de expresiones recolectadas. La tabla 5 presenta los datos principales de este corpus.

```
Por otra parte me parece muy
mejor que otros pues por lo que he visto
no tiene por que ser la misma que
por favor intenta de nuevo
ahora bien lo que no se si se puede hacer es
```

Tabla 4. Fragmento del corpus WebDIME

27,224,579	total de instancias de formas léxicas
1,129	total formas léxicas
4,520,513	total de secuencias
6	promedio de palabras por secuencia

Tabla 5. Datos principales del corpus WebDIME

3. Análisis comparativo de los corpus

Es claro que las expresiones y términos usados en diálogos reales difieren de los usados en los textos escritos. Por ejemplo, es obvio suponer que la frecuencia de pronombres (o verbos conjugados) en primera y segunda persona no son iguales en un diálogo entre dos personas y un documento escrito. El propósito de este estudio es identificar claramente los términos cuyas proporciones de ocurrencias entre ambos corpus difieren considerablemente, ya sea porque estos términos se encuentran sobre o subrepresentados.

La metodología propuesta para el análisis comparativo del léxico de los corpus considera los siguientes procesos:

1. *Preprocesamiento de los corpus*, que se enfoca principalmente en la generación de sus índices, y en la construcción de sus distribuciones de términos.

2. *Comparación léxica de los corpus*, que considera la medición de la disimilitud entre las distribuciones de probabilidad de los términos en los corpus, y la identificación de las palabras críticas.

A continuación se describen ambos procesos.

3.1 Preprocesamiento de los corpus

Esta primera etapa de procesamiento se refiere a la creación de un *índice de los corpus*. Dicho índice indica los términos usados en cada corpus y sus frecuencias de ocurrencia. Su representación es mediante un *archivo invertido* [Kowalski, 97], es decir, una estructura de datos que consta de un diccionario y una lista invertida instrumentados a través de tablas de *hash*. En el diccionario se almacenan todos los términos extraídos, junto con su frecuencia total de ocurrencia. En la lista invertida se almacena, para cada término, una lista dinámica de las colecciones (podrían ser más de dos) en los que el término fue encontrado y la frecuencia de ocurrencia en cada una de ellas.

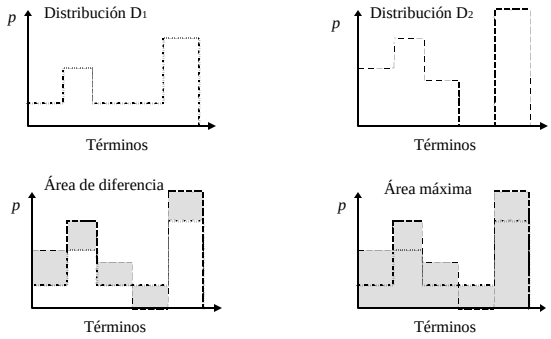

Figura 1. Comparación de las distribuciones de probabilidad

A partir del índice construido, una frecuencia f_k^i es asignada a cada uno de los términos. Esta frecuencia indica el número de ocurrencias del término k en cada corpus (en nuestro caso, que sólo son dos corpus i ={1,2}). Con base en estas frecuencias se construye una distribución de probabilidad $D_i = \{p_k^i\}$ de los términos en el corpus i, donde $p_k^i = f_k^i \Big/ \sum_{j=1}^{n} f_j^i$ expresa la probabilidad de ocurrencia del término k en el corpus i, y n indica el número de términos existentes en el índice.

3.2 Comparación léxica de los corpus

3.2.1 Comparación de las distribuciones de probabilidad

Para medir la diferencia léxica entre dos corpus se comparan sus distribuciones de probabilidad D_i = { p_k^i } para i ={1,2}. El propósito es medir la diferencia absoluta entre los corpus, sin considerar a ninguno de ellos como punto de referencia, se propuso la medida C_d para comparar las distribuciones. Esta medida se expresa como el cociente del área de diferencia entre el área máxima de las distribuciones de probabilidad (ver figura 1). Esta medida refleja la diferencia entre los corpus y no se ve afectada por las diferencias relativas de cada uno de los términos. Para una revisión más detallada de estos conceptos véase [Montes-y-Goméz et al., 2001].

$C_d = \dfrac{A_d}{A_m}$ coeficiente de diferencia:

$A_d = \sum_{k=1}^{n} d_k$ área de diferencias

$A_m = \sum_{k=1}^{n} \max\left(p_k^1,\ p_k^2\right)$ área máxima

$d_k = \left| p_k^1 - p_k^2 \right|$ diferencias de términos

Si el coeficiente de diferencia entre las dos distribuciones de probabilidad es cercano a 1, entonces existe una diferencia considerable entre las ocurrencias de los términos de los dos corpus. Si por el contrario, el coeficiente de diferencia es cercano a 0, entonces se puede concluir que los dos corpus son léxicamente similares.

3.2.3 Identificación de las palabras críticas

La diferencia global entre dos corpus se origina por la diferencia abrupta de varios términos individuales. Estos términos pueden definirse como aquellos que presentan una diferencia notablemente mayor que la diferencia promedio. Entonces, considerando que d_m es el valor típico de d_k, y d_s es la medida de dispersión de la distribución, los términos para los que $d_k > d_m + \left(C \times d_s\right)$ pueden considerarse los más diferentes. Ajustando la constante C se puede determinar el criterio usado para identificar un término como *crítico*.

$d_\mu = \dfrac{1}{n}\sum_{k=1}^{n} d_k$ diferencia promedio

$d_s = \sqrt{\dfrac{1}{n}\sum_{k=1}^{n}\left(d_k - d_\mu\right)^2}$ desviación estándar de las diferencias

Estadísticas	Valor
Vocabulario para DIME y WebDIME	1129
Coeficiente de Diferencia (C_d)	0.717
Área de Diferencia (A_d)	1.11767
Área Máxima (A_m)	1.55883
Criterio para identificar factores de cambio por promedio.	
Cambio Promedio (d_m)	0.00098
Desviación Estándar del cambio (d_s)	0.00337

Tabla 6. Resultado de las estadísticas realizadas

4. Resultados

Como se mencionó en párrafos anteriores el estudio tiene por objetivo identificar claramente los términos cuyas frecuencias en ambos corpus se alejan considerablemente, es decir, el conjunto de *palabras críticas*. La tabla 6 muestra un concentrado de los resultados del presente estudio. Los puntos más importantes a remarcar son:

- *El coeficiente de diferencia*. A través de este coeficiente podemos comprobar una diferencia importante entre las proporciones de ocurrencias de los términos en ambos corpus. Es importante recordar que mientras más cercano esté el coeficiente a 1 la diferencia es mayor. En este caso el coeficiente de diferencia es de 0.71.

- *Identificación de grupos de palabras críticas*. El conjunto de palabras críticas obtenido representa el 4% de todos los términos (véase la tabla 7). La figura 2 muestra la diferencia en las proporciones de las palabras críticas. De dicho conjunto de palabras son de particular importancia los términos subrepresentados, es decir, aquellos términos que están presentes de forma abundante en el corpus DIME y, que por el contrario, su frecuencia en el corpus WebDIME es relativamente pequeña.

Dentro de las palabras críticas es interesante notar cuatro subgrupos:

- *Vocablos específicos del dominio*. Términos como refrigerador, alacena o estufa están mal representados. Esto es de particular cuidado ya que dentro de la aplicación dichos términos son muy comunes. Cabe hacer notar que palabras como mesa o silla no están dentro del conjunto de palabras críticas.

- *Deícticos gramaticales*. Dentro del conjunto de palabras críticas también encontramos deícticos gramaticales, es decir, vocablos que acompañados con un gesto sirven para identificar elementos del discurso, p.e. ahí, aquí, éste. La mala representación de estos términos es decisiva dada la naturaleza multimodal de la aplicación en cuestión.

- *Verbos involucrados en expresiones de cortesía*. Las formas léxicas de verbos como: poder o querer están subrepresentados. Este resultado es de esperarse dado que las formas léxicas de estos verbos son abundantes en el habla oral espontánea y son casi totalmente ausentes en el lenguaje escrito.

- *Palabras vacías*. Del otro extremo del panorama encontramos un conjunto de palabras sobrerepresentadas. Dichas palabras generalmente son artículos y preposiciones.

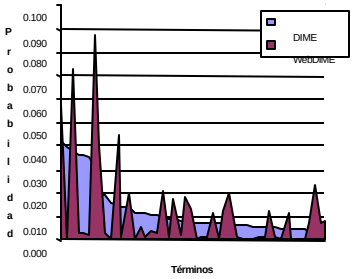

Figura 2. Proporciones de las palabras críticas

Palabras subrepresentadas			*Palabras sobrerepresentadas*	
ahí	está	ponga	como	los
ahora	este	puedes	con	no
alacena	éste	quieres	de	para
alacenas	estufa	quiero	del	por
aquí	fregadero	refrigerador	el	que
así	hacia	sí	en	se
bien	mueble	tenemos	es	si
bueno	okey	vamos	la	su
dónde	pared	ver	las	una
esta	poner		lo	

Tabla 7. Conjunto de palabras críticas

5. Conclusiones y perspectivas

Este trabajo presentó una metodología para identificar los puntos débiles a nivel léxico de un corpus orientado al entrenamiento de modelos de lenguaje. Dicha metodología fue puesta en práctica usando las colecciones DIME y WebDIME. Ello permitió identificar un conjunto de 48 palabras críticas. Por supuesto, la falta de muestras del uso de estas palabras afecta el rendimiento del modelo de lenguaje calculado al usar WebDIME. Esto se debe a dos factores principales: (i) el daño en la cobertura del modelo de lenguaje se extiende a todos los contextos de uso de dichas palabras; y (ii) como puede observarse en la tabla 7, palabras propias del dominio de aplicación están mal representadas en el corpus WebDIME.

Evidentemente, la identificación de los puntos débiles es sólo el primer paso. Etapas posteriores serán necesarias para enriquecer adecuadamente el corpus en cuestión, y así entrenar modelos de lenguaje más robustos. Específicamente para el caso del corpus WebDIME y a partir de los resultados hallados, se ha pensado en aplicar dos tipos de enriquecimiento. El primero orientado a incrementar la presencia de términos del domino. Para ello se enriquecerá el corpus con textos seleccionados cuya temática sea el diseño de cocinas. Métodos como Clasitex [Gelbukh et al., 99] se usarán para detectar el dominio y así colectar los textos de interés. Por otro lado, también será necesario enriquecer el corpus WebDIME con algunos fenómenos propios del habla oral (deícticos gramaticales y expresiones de cortesía). En este caso, el mismo corpus DIME se usará como fuente de datos. Sin embargo, ello implica definir un método para alcanzar porcentajes de ocurrencia pertinentes para dichos términos, sin alterar las proporciones de los términos bien representados.

Otros puntos a explorar en trabajos futuros son: el análisis comparativo a nivel sintáctico; el enriquecimiento del vocabulario con las palabras más frecuentes del español; y el enriquecimiento del corpus con documentos extraídos de foros de discusión.

Agradecimientos

El presente trabajo se realizó con el apoyo parcial del CONACyT (proyecto 31128-A), del Laboratorio de Tecnologías del Lenguaje del INAOE, y dentro del marco de cooperación bilateral del Laboratorio Franco-Mexicano de Informática (LAFMI).

Referencias

[Bernsen et al., 98] Bernsen, N., Dybkjaer, H. & Dybkjaer, L. *Designing Interactive Speech Systems. From First Ideas to User Testing*. Springer-Verlag. 1998.

[Gelbukh et al., 02] Gelbukh, A., Sidorov, G. & Chanona, L. *Compilation of a Spanish Representative Corpus*. Computational Linguistics and Intelligent Text Processing, International Conference CICLing02, Lecture Notes in Computer Science 2276, Springer. 2002.

[Gelbukh et al., 99] Gelbukh, A., Sidorov, G. & Guzman-Arenas, A. *Use of a weighted topic hierarchy for document classification*. In Václav Matoušek *et al*. (Eds.). Text, Speech and Dialogue. Proc. 2nd International Workshop TSD-99, Plzen, Czech Republic, September 13-17, 1999. Lecture Notes in Artificial Intelligence 1692, ISSN 0302-9743, ISBN 3-540-66494-7, Springer-Verlag, pp. 130–135.

[Jurafsky & Martin, 00] Jurafsky, D. & Martin, J. *Speech and Language Processing*. Prentice Hall. 2000.

[Kowalski, 97] Kowalski, G. *Information Retrieval Systems: Theory and implementation*, Kluwer Academic Publishers, 1997.

[Montes-y-Gómez et al., 01] Montes y Gómez, M. Gelbukh, A. & López, *A. Mining the News: Trends Associations and Deviations*. Computación y Sistemas. Vol. 5. No. 1. pp 14-24. ISSN 1405-5546. CIC-IPN 2001.

[Pineda et al., 02] Pineda, L. A., Massé, A., Meza, I., Salas, M., Schwarz, E., Uraga, E. & Villaseñor, L. *The DIME Project*. MICAI 2002, Carlos Coello (eds.) Lecture Notes in Artificial Intelligence 2313, pp. 166-175. Springer-Verlag. 2002.

[Vaufreydaz et al., 99] Vaufreydaz, D., Akbar, M. & Rouillard, J. *Internet Documents : A Rich Source for Spoken Language Modeling*. Automatic Speech Recognition and Understanding (ASRU`99), Keystone, Colorado (USA), p.277-280

[Villaseñor et al., 01] Villaseñor, L., Massé, A. & Pineda, L.A. *The DIME corpus*. ENC01, 3er Encuentro Internacional de Ciencias de la Computación, Aguascalientes, México. SMCC-INEGI 2001.