Evaluación de características lingüísticas (sintácticas, morfológicas y semánticas) de frases orientada al análisis de dificultad en Recuperación de Información

Alejandro Bellogin Kouki

16 de junio de 2008

1. Introducción

En este trabajo se van a utilizar técnicas de Procesamiento de Lenguaje Natural (PLN) para calcular una serie de *predictores de dificultad*, de manera que, dada una frase, se obtenga un valor numérico que indique el grado de dificultad de la misma. Las técnicas de PLN utilizadas comprenden tanto el nivel sintáctico, como morfólogico y semántico de la frase. La salida de este análisis puede ser útil para otras áreas además del área en la que nos centraremos: la Recuperación de Información (ver sección 5.1).

2. Análisis de dificultad

En los últimos años, el interés por el análisis de dificultad en el campo de Recuperación de Información ha aumentado bastante, debido a que puede ser útil para mejorar el rendimiento de los sistemas de búsqueda. Los trabajos que más aceptación han tenido han sido los que utilizan implícitamente métodos de recuperación de información (realizan búsquedas con la consulta a estudiar y tienen en cuenta los documentos que devuelve), no obstante, también se han estudiado técnicas que analizan morfológica, sintáctica y semánticamente la consulta. Estos últimos requieren menos procesamiento, ya que no se realiza la búsqueda propiamente dicha; permitirían, por tanto, modificar los algoritmos de búsqueda en función de la información extraída de la consulta.

En las siguientes secciones se explican los distintos predictores que se engloban en estas dos grandes categorías.

2.1. Predictores lingüísticos

Tomando como punto de partida los predictores expuestos en [6], se han extraído las siguientes características de cada sentencia:

- Número de palabras
- Longitud media de las palabras

- Número de nombres propios
- Número de pronombres
- Número de acrónimos
- Número de conjunciones
- Número de preposiciones
- Número de significados de cada palabra (polisemia): se calcula como el número de *synsets* devuelto por Wordnet.
- Número de polisemia normalizado (medio)
- Profundidad sintáctica: máximo número de componentes sintácticos anidados
- Anchura de los enlaces sintácticos: es el cociente entre la distancia entre cada enlace sintáctico (medida por el número de palabras que los separan) y el número de enlaces sintácticos

Un ejemplo de los dos últimos predictores se puede ver en la imagen 1.

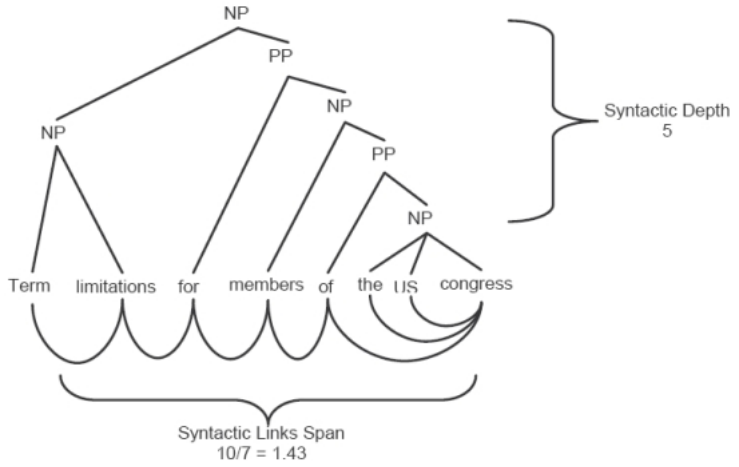

Figura 1: Ejemplos de valores de profundidad sintáctica (*syntactic depth*) y anchura de los enlaces sintácticos (*syntactic links span*).

Debido a que no se encontraron herramientas equivalentes a algunas de las utilizadas en el artículo, no se han podido implementar los siguientes indicadores:

- Número medio de morfemas por palabras
- Número medio de sufijos por palabra

No obstante, se han añadido los siguientes:

- Número de hipónimos[1]: se calcula con Wordnet
- Número medio de hipónimos

[1]Palabra cuyo significado es más específico que el de otra en la que está englobada

2.2. Predictores no lingüísticos

En [3] y [4] se muestran varios predictores de dificultad, enmarcados en el área de la Recuperación de Información, que no utilizan características lingüísticas de las frases sino probabilidades de relevancia, calculadas como frecuencias de aparición:

- Claridad: es una medida de similitud entre los modelos de lenguaje asociados a la consulta y a una colección, ya que se entiende que cuanto más se parezca el modelo de lenguaje de la consulta al de la colección, más ambigua (menos clara) es. Se calcula mediante la siguiente fórmula [3]:

$$\begin{aligned} P(w|Q) &= \sum_{D \in R} P(w|D)P(D|Q), \quad P(Q|D) = \prod_{q \in Q} P(q|D) \\ P(w|D) &= \lambda P_{ml}(w|D) + (1-\lambda)P_{coll}(w) \\ \text{Claridad} &= \sum_{w \in V} P(w|Q) \log_2 \frac{P(w|Q)}{P_{coll}(w)} \end{aligned}$$

  donde w es cualquier término, Q es la consulta, D es un documento, R es el conjunto de documentos que contiene al menos un término de la consulta, $P_{ml}(w|D)$ esl a frecuencia relativa del término w en el documento D, $P_{coll}(w)$ es la frecuencia relativa del término en la colección entera, λ es un parámetro (en su trabajo, 0,6) y V es todo el vocabulario.

  Esta medida es la entropía relativa entre los dos modelos de lenguaje. Se ha demostrado que está muy relacionada con el rendimiento de un sistema de Recuperación de Información.

- Claridad simplificada (SCS, de *simplified clarity score*): dado que el cálculo de la claridad es muy pesado, en [4] proponen una versión simplificada del mismo:

$$\begin{aligned} SCS &= \sum_{Q} P_{ml}(w|Q) \log_2 \frac{P_{ml}(w|Q)}{P_{coll}(w)} \\ P_{ml}(w|Q) &= \frac{qtf}{ql} \end{aligned}$$

  donde qtf es el número de ocurrencias del término w en la consulta y ql es la longitud de la consulta.

  Está demostrado que este predictor está fuertemente correlacionado con el rendimiento en consultas cortas.

- Alcance de la consulta (QS, de *query scope*): en [4] utilizan el siguiente indicador como una medidad de la generalidad/especificidad de una consulta, utilizando para ello el tamaño del conjunto de respuesta para una determinada consulta; de esta manera:

$$QS = -\log\left(n_Q/N\right)$$

  donde n_Q es el número de documentos que contienen al menos un término de la consulta y N es el número de documentos en toda la colección.

  La correlación entre este predictor y la precisión promedio ha demostrado ser significativa sólo para consultas cortas (ya que en otro caso, se estabiliza).

3. Herramientas utilizadas

Con el objetivo de analizar el texto y obtener los indicadores que se acaban de describir, se han buscado herramientas (gratuitas) que permitieran hacer dicho análisis. Finalmente, se eligieron las herramientas que describiremos a continuación:

nltk Grupo de programas escritos en Python y que permiten varias tareas de procesamiento de lenguaje natural. Iba a utilizarse para crear un árbol sintáctico dado una frase, pero debido a la necesidad de definir una gramática (y a la fuerte dependencia de los resultados con respecto a esta gramática) se descartó.

Wraetlic Descritas en [1], son un conjunto de herramientas que permiten desde marcar el texto con etiquetas POS hasta identificar sintagmas. En un principio, se iban a utilizar sólo para algunas partes del análisis morfológico y sintáctico, pero han terminado resultando una parte esencial del trabajo.

Además de herramientas de Procesamiento de Lenguaje Natural, se ha utilizado la base de datos de Wordnet y otras herramientas, una matemática y otra propia del campo de Recuperación de Información:

Matlab Entorno de desarrollo y programación para programación numérica. PErmite manipular matrices de manera muy sencilla y numerosas operaciones estadísticas y gráficas.

Terrier Esta herramienta provee funcionalidades para indexar y recuperar documentos, utilizando distintos algoritmos y permitiendo adaptarlo a las necesidades de cada aplicación de una manera muy sencilla. Está escrito en Java.

Además de las herramientas mencionadas, se estudiaron y probaron otras, comprobando el funcionamiento y características de las mismas para saber si iban a ser útiles en el trabajo o no. Las herramientas revisadas se pueden ver en la tabla 1.

Como se puede observar, al final no se han utilizado muchas de las herramientas revisadas; esto ha sido así ya que se ha conseguido hacer la mayor parte del análisis requerido usando Wraetlic.

4. Resultados

Con el objetivo de obtener resultados comparables con el resto de la literatura, se ha seguido la siguiente metodología (igual que la de los autores de [5, 6]):

1. Se elige un conjunto de consultas descritas en alguna tarea de TREC[16] (*Text REtrieval Conference*). Esto se hace ya que TREC es una conferencia que se celebra anualmente, donde se proponen *tracks* (áreas de interés en donde se definen determinadas tareas de recuperación de información). Cada track lleva asociado un conjunto de *topics* (preguntas o consultas), un conjunto de documentos (sobre los que se lanzan las preguntas) y las respuestas correctas (juicios de relevancia). Según este último elemento se obtiene un ranking entre los sistemas que participan en la conferencia.

[16] http://trec.nist.gov

Nombre	Características	Utilizada?
Bios[2]	Analizador sintáctico-semántico	No
Celex[3]	Base de datos con descomposición morfológica	(No disponible)
Gate[4]	Tokenización, etiquetas POS, partición de frases, reconocimiento de entidades, resolución de coreferencias	No
jBeaver[5]	Análisis de generación de analizadores de dependencias	No
jTextPro[6]	Partición de frases, tokenización, etiquetas POS, identificación de sintagmas	No
Matlab[7]	Herramienta utilizada para análisis estadístico	Sí
nltk[8]	Procesamiento de lenguaje natural estadístico y simbólico (etiquetas POS, identificación de sintagmas), clasificador de textos	Sí
tagHelper[9]	Analisis de textos en inglés, alemán, español y chino	No
Terrier[10]	Motor de búsqueda, indexador	Sí
Tree-tager[11]	Etiquetas POS, tokenización	Sí
Spear[12]	Parser de dependencias	No
SwiRL[13]	Etiquetador semántico, análisis sintáctico	No
Syntex	Análisis sintáctico	(No encontrado)
Wordnet[14]	Base de datos léxica	Sí
Wraetlic[15]	Etiquetas POS, tokenización, partición de frases, identificación de sintagmas, desambiguación, reconocimiento de entidades	Sí

Cuadro 1: Herramientas revisadas

2. Se utilizan los resultados de los sistemas que participan en TREC para calcular el rendimiento esperado de cada consulta. Esto se realiza mediante una medida llamada Average Precision (AP), cuya fórmula es

$$AP = \frac{\text{documentos relevantes respondidos}}{\text{documentos respondidos}}$$

Se realiza el promedio de las precisiones medias de cada sistema y se obtiene un valor para cada consulta.

3. Se ejecutan los predictores a analizar con cada consulta (en particular, se utiliza sólo el título de cada una, dado que los otros campos (descripción, narración) no se parecen a una consulta real de usuario por ser muy extensos).

4. Finalmente, se estudia la correlación entre los valores devueltos por cada predictor y el promedio de las precisiones medias de cada una de las consultas. De las correlaciones calculadas, son relevantes las estadísticamente significativas. En este caso, se han utilizado tres coeficientes de correlación distintos: ρ de Spearman, τ de Kendall y r de Pearson (Spearman es un caso particular de Pearson, Kendall es un estadístico no paramétrico); como se verá a continuación, no se obtienen los mismos resultados con los tres análisis.

Los resultados obtenidos en este caso, usando las consultas de TREC 8, TREC 9 y TREC 2001 (150 consultas en total), se encuentran en las tablas 2 y 3.

Hay que destacar, que el hecho de que los textos de prueba estén en inglés es una ventaja, ya que la mayoría de las herramientas (sólo) soportan este lenguaje.

Consultas	**Pearson**	**Spearman**	**Kendall**
Todas	Nombres propios, hiponimia, polisemia, polisemia normalizada	Nombres propios, polisemia, polisemia normalizada	Nombres propios, polisemia, polisemia normalizada
TREC 8	Nombres propios, profundidad sintáctica	Nombres propios, profundidad sintáctica	Nombres propios, profundidad sintáctica
TREC 9	Nombres propios, hiponimia, polisemia normalizada	Hiponimia normalizada, polisemia normalizada	Hiponimia normalizada, polisemia normalizada
TREC 2001	Acrónimos	Acrónimos	Acrónimos

Cuadro 2: Predictores lingüísticos encontrados correlacionados de manera estadísticamente significativa con la precisión promedio

Consultas	**Pearson**	**Spearman**	**Kendall**
Todas	SCS, claridad	SCS, claridad	SCS, claridad
TREC 8	Scope, SCS	Scope, SCS, claridad	Scope, SCS, claridad
TREC 9		SCS	SCS
TREC 2001	Claridad	Claridad	Claridad

Cuadro 3: Predictores no lingüísticos encontrados correlacionados de manera estadísticamente significativa con la precisión promedio

Se puede observar que, al igual que [6], los predictores "sencillos" no aparecen en las tablas, sino que son los predictores complejos, como el número de hipónimos, de polisemia o la profundidad sintáctica, los que aparecen mayor número de veces como significativos. También se observa que las consultas tienen características distintas en cada conferencia, ya que no se obtienen los mismos resultados, por ejemplo, la precisión de las consultas de TREC 8 son las únicas que están correlacionadas con la profundidad sintáctica, mientras que Pearson aplicado a TREC 9 no encuentra ningún predictor no lingüístico estadísticamente significativo.

5. Trabajo futuro

Los resultados mostrados se centran en la correlación entre los resultados de los distintos predictores y la medida de dificultad definida en este caso (promedio de las precisiones medias). En esta situación, no necesitamos saber el rango de valores de cada uno de los predictores, ni establecer fronteras para definir cuándo una frase es difícil y cuándo no; en general, sería muy útil conocer estos datos o tener un conocimiento aproximado de ellos. De hecho, una posible aplicación una vez se conocieran estos

umbrales, sería crear un predictor a partir de un algoritmo genético que combinara todos los que hemos descrito.

Estos predictores tienen un problema en su propia definición: ¿qué es la dificultad? La respuesta a esta pregunta depende del campo de aplicación, y un análisis de una respuesta general sería un trabajo muy interesante a desarrollar.

En la siguiente sección describimos distintas aplicaciones que pueden tener estos predictores, indicando algunos campos en los que se utiliza en la actualidad.

5.1. Posibles campos de aplicación

Identificar la dificultad inherente a una sentencia de manera automática puede ser útil en muchos campos. Como ya se ha visto, en Recuperación de Información se puede utilizar para descubrir cuánto de difícil o de fácil es una consulta (lo cual está relacionado con el grado de ambigüedad de la misma [3]). Esta información ayudaría a las siguientes tareas:

- Ayudar al usuario a formular mejor las consultas, ya que se podría avisar con antelación a la persona de que es muy probable que su consulta sea malinterpretada por el sistema, existiendo la posibilidad de sugerir distintas consultas a partir de una dada [2].
- Adaptar el algoritmo de búsqueda en función del valor de complejidad esperado para cada consulta.
- En sistemas de metabuscadores (un buscador que aglutina el resultado de otros buscadores) se puede utilizar el valor de ambigüedad de una consulta en relación con cada buscador para ponderar el ranking final (los resultados del sistema para el que una consulta es menos ambigua estarían más acotados y, por tanto, serían mejores).
- Ayudar al administrador del sistema a descubrir por qué el buscador responde mal a determinadas consultas y si esto está relacionado con la complejidad de las consultas o no.

En sistemas Hipermedia Adaptativos para la educación normalmente existe una parte que sirve para evaluar al estudiante, para esta parte el profesor elige una pregunta que espera sea contestada satisfactoriamente por el estudiante. La tendencia en la actualidad es realizar esta evaluación de manera automática, es decir, que se puedan evaluar respuestas libres (dadas por el estudiante) y compararlas con una respuesta prefijada por el profesor. El análisis de la dificultad de una sentencia se podría utilizar en la fase de creación del curso para dar al profesor una medida de cuánto de fácil sería responder dicha pregunta con el material del curso, y, con esta información, el profesor podría decidir si la pregunta es adecuada para el nivel y/o la fase del curso en el que se encuentra.

Otra posible aplicación de este análisis lo encontraríamos en los sistemas de diálogo, los cuales podrían solicitar una reformulación de la consulta cuando descubrieran una frase difícil.

Un inconveniente de las técnicas lingüísticas que hemos presentado es que necesitan ser aplicadas sobre lenguaje natural, no como las técnicas no lingüísticas. Esto provoca que no sean aplicables cuando se quiera extrapolar medidas de dificultad a otros ámbitos (averiguar cuánto de “ambiguo” es un usuario dentro de un conjunto

de usuarios a partir de su conjunto de preferencias o un conjunto de términos que lo definan).

6. Conclusión

En este trabajo se han mostrado diversas técnicas que permiten analizar la dificultad de una sentencia, algunas empleando Procesamiento de Lenguaje Natural y otras, Teoría de la Información. También se ha visto que algunas de ellas están relacionadas con la dificultad de una frase, si entendemos que la *dificultad* está relacionada con el rendimiento de un sistema de Recuperación de Información. No obstante, estas técnicas no deberían estar relegadas sólo al campo de la Recuperación de Información.

Otro resultado de este trabajo ha sido la experiencia obtenida al haber probado varias herramientas de Procesamiento de Lenguaje Natural.

Referencias

[1] Enrique Alfonseca, Antonio Moreno-Sandoval, José M. Guirao, and Maria Ruiz-Casado. The wraetlic nlp suite. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genova, Italy, May 2006.

[2] James Allan and Hema Raghavan. Using part-of-speech patterns to reduce query ambiguity. In *25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 307–314, 2002.

[3] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Predicting query performance. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306, New York, NY, USA, 2002. ACM.

[4] Ben He and Iadh Ounis. Inferring query performance using pre-retrieval predictors. In *String Processing and Information Retrieval, SPIRE 2004*, pages 43–54, 2004.

[5] Josiane Mothe and Ludovic Tanguy. Linguistic analysis of users'queries: towards an adaptive information retrieval system.

[6] Josiane Mothe and Ludovic Tanguy. Linguistic features to predict query difficulty. In *Predicting Query Difficulty - Methods and Applications, SIGIR 2005*, 2005.