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

Outline

Introducción

Qué se va a hacer

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.

Introducción

Qué se va a hacer

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.

Qué se entiende por dificultad

La dificultad de una consulta está relacionada con su precisión promedio (en RI):

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

En general, esta definición depende del campo en el que se esté.

Outline

Predictores utilizados

Predictores lingüísticos

Utilizan características morfológicas, sintácticas y semánticas de la frase a analizar

Predictores utilizados

Predictores lingüísticos

Utilizan características morfológicas, sintácticas y semánticas de la frase a analizar

Predictores no lingüísticos

Utilizan métodos de recuperación de información (búsquedas con la consulta a estudiar teniendo en cuenta los documentos que devuelve) y probabilidades de relevancia, calculadas como frecuencias de aparición

Predictores lingüísticos I

Se han extraído las siguientes características de cada sentencia [6]:

- 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

Predictores lingüísticos II

Ejemplo de los dos últimos predictores:

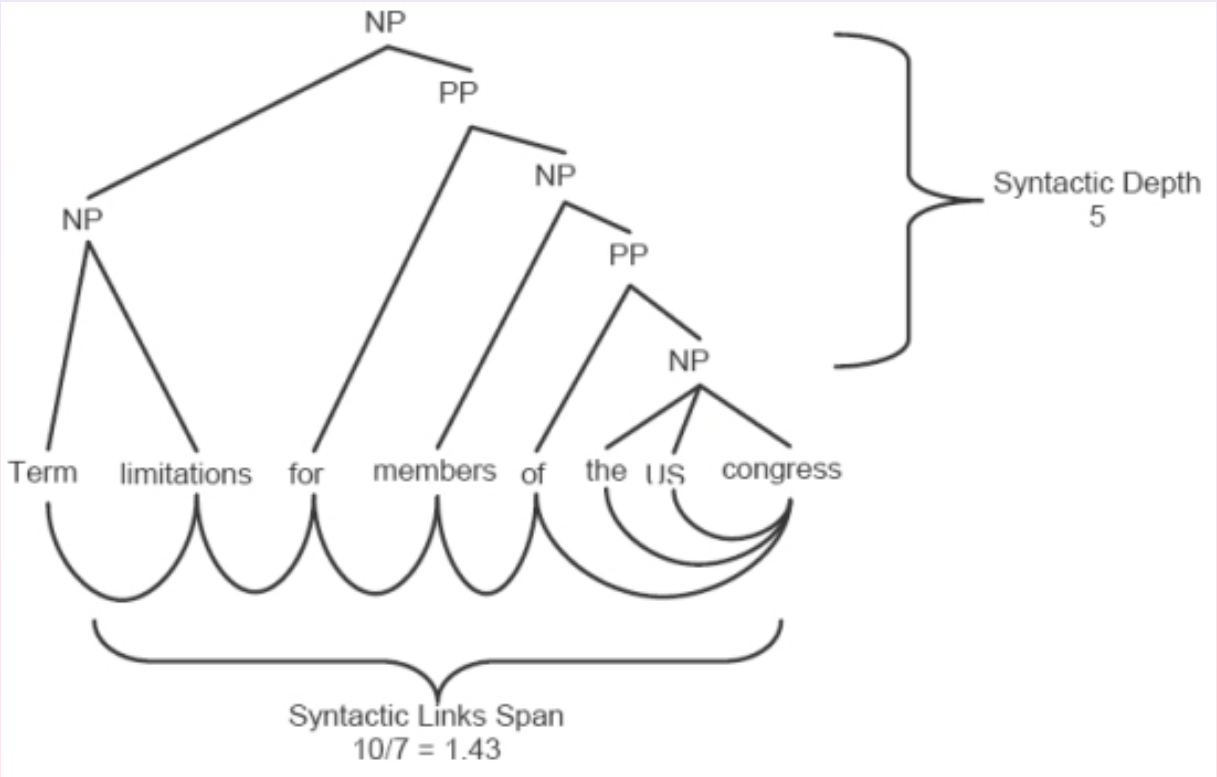

Figura: Ejemplos de valores de profundidad sintáctica (*syntactic depth*) y anchura de los enlaces sintácticos (*syntactic links span*).

Predictores lingüísticos III

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

Predictores no lingüísticos I

En [3] y [4] se muestran varios predictores de dificultad:

- 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.

Predictores no lingüísticos II

- 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:

$$SCS = \sum_{Q} P_{ml}(w|Q) \log_2 \frac{P_{ml}(w|Q)}{P_{coll}(w)}$$

$$P_{ml}(w|Q) = \frac{qtf}{ql}$$

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).

Outline

Herramientas utilizadas

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: Herramientas revisadas

Outline

Metodología

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.
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).
   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.

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

Resultados I

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

Resultados I con correlaciones

Queries	Pearson	Spearman	Kendall
All	Proper nouns (0,2305), hyponymy (−0,1808), polysemy (−0,1933), normalized polysemy (−0,2799)	Proper nouns (0,2103), polysemy (−0,2089), normalized polysemy (−0,2506)	Proper nouns (0,1726), polysemy (−0,1414), normalized polysemy (−0,1685)
TREC 8	Proper nouns (0,2857), syntactic depth (−0,1201)	Proper nouns (0,3360), syntactic depth (−0,0275)	Proper nouns (0,2772), syntactic depth (−0,0211)
TREC 9	Proper nouns (0,2978), hyponymy (−0,3084), normalized polysemy (−0,3218)	Normalized polysemy (−0,3445), normalized hyponymy (−0,3099)	Normalized hyponymy (−0,2177), normalized polysemy (−0,2276)
TREC 2001	Acronyms (0,3626)	Acronyms (0,2814)	Acronyms (0,2320)

Cuadro: Linguistic features found statistically significant correlated with average precision (correlation in brackets, the greater absolute value, the more dependance between variables)

Resultados II

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

Resultados II con correlaciones

Queries	Pearson	Spearman	Kendall
All	SCS (0,2615), clarity (−0,2154)	SCS (0,3519), clarity (−0,3005)	SCS (0,2361), clarity (−0,2003)
TREC 8	Scope (0,4771), SCS (0,6037)	Scope (0,3248), SCS (0,4919), clarity (−0,3268)	Scope (0,2640), SCS (0,3339), clarity (−0,2327)
TREC 9		SCS (0,4402)	SCS (0,3011)
TREC 2001	Clarity (−0,4822)	Clarity (−0,4452)	Clarity (−0,3004)

Cuadro: Non-linguistic features found statistically significant correlated with average precision (correlation in brackets, the greater absolute value, the more dependance between variables)

Outline

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.
- Campos de aplicación:
  - Recuperación de Información
  - Sistemas Hipermedia Adaptativos para la educación
  - Sistemas de diálogo

Conclusiones

- 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.
- 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.
- Estas técnicas no deberían estar relegadas sólo al campo de la Recuperación de Información.

Outline


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.


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.


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.

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

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

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