Experimentos con Métodos de Extracción de la Idea Principal de un Texto sobre una Colección de Noticias Periodísticas en Español

Tolosa, G. H..; Peri, Jorge A. y Bordignon, Fernando R. A
{tolosoft, jperi, bordi}@unlu.edu.ar
Universidad Nacional de Luján
Departamento de Ciencias Básicas
Laboratorio de Redes de Datos

Resumen

Este trabajo tiene por objetivo evaluar distintas técnicas de selección de la oración que mejor representa la idea principal (*gist*) de un texto corto (noticia de prensa) en español. Se utilizaron dos métodos clásicos de la literatura y dos propuestas extra, basadas en la estructura de los documentos. Para la evaluación se tomó el proceso de clasificación para determinar la capacidad del extracto para "representar" el contenido de la noticia y su categoría.

Los métodos evaluados obtienen un buen comportamiento, superando – en algunos casos – el 90% de eficiencia respecto del experimento utilizando la noticia completa. Complementariamente, se realizaron experiencias utilizando combinaciones de los mejores métodos, logrando un incremento de la eficiencia de un 5%.

Palabras clave: resumen automático, extracción de la idea principal.

Workshop de Ingeniería de Software y Bases de Datos (WISBD)

Experimentos con Métodos de Extracción de la Idea Principal de un Texto sobre una Colección de Noticias Periodísticas en Español

Tolosa, G. H..; Peri, Jorge A. y Bordignon, Fernando R. A
{tolosoft, jperi, bordi}@unlu.edu.ar
Universidad Nacional de Luján
Departamento de Ciencias Básicas
Laboratorio de Redes de Datos

Resumen

Este trabajo tiene por objetivo evaluar distintas técnicas de selección de la oración que mejor representa la idea principal (*gist*) de un texto corto (noticia de prensa) en español. Se utilizaron dos métodos clásicos de la literatura y dos propuestas extras, basadas en la estructura de los documentos. Para la evaluación se tomó el proceso de clasificación para determinar la capacidad del extracto para "representar" el contenido de la noticia y su categoría.

Los métodos evaluados obtienen un buen comportamiento, superando – en algunos casos – el 90% de eficiencia respecto del experimento utilizando la noticia completa. Complementariamente, se realizaron experiencias utilizando combinaciones de los mejores métodos, logrando un incremento de la eficiencia de un 5%.

1. Introducción

El tratamiento automático de la información facilita a los usuarios la manipulación, evaluación y utilización de grandes cantidades de documentos, tarea que – mediante técnicas completamente manuales – no se podría realizar. Esta situación se encuentra afectada por la alta disponibilidad de documentos en Internet y la existencia de fuentes de información "en línea", como por ejemplo las agencias de noticias y periódicos digitales.

Una de las técnicas que intenta brindar soluciones a algunos de estos inconvenientes es el resumen automático de textos (summarization), proceso por el cual se intenta identificar la información sustancial de un texto para generar una versión abreviada [7] [8].

Los resúmenes permiten caracterizar el contenido de un texto en una cantidad limitada de palabras u oraciones, de acuerdo a la necesidad y a la aplicación, modificando el procesamiento posterior que se requiere. Los ambientes de aplicación de estas técnicas son amplios y – en principio – de cualquier texto se puede extraer un resumen: noticias, artículos científicos, tesis, libros. Desde el punto de vista de la utilidad, los resúmenes permiten reducir el tiempo de procesamiento, los requisitos de almacenamiento y el tráfico de datos en las transferencias electrónicas.

Un concepto relacionado con el de resumen de un texto es el de “idea principal” (denominada gist). Para Pardo [9] todo texto posee una idea principal y es posible identificar aquella oración que mejor la expresa. Se puede plantear – entonces – una analogía entre esta oración y un resumen mínimo de un texto, de solo una oración.

La idea principal de un texto es un atributo importante que puede ser útil para diversas aplicaciones, tales como resúmenes automáticos, clasificación de textos y obtención de nuevos metadatos para documentos web. En este último caso, esta información se puede asociar al rankeo de los documentos en un sistema de recuperación de información. De manera general, el proceso de extracción de la idea principal puede ser visto como uno de selección de características de un documento (*feature selection*), las cuales se utilizarán en tareas posteriores.

Este trabajo tiene por objetivo evaluar distintas técnicas de selección de la oración que mejor representa la idea principal de un texto corto (noticia de prensa) en español. Para ello, se toma como criterio de evaluación la clasificación. Se entiende por clasificación al proceso de asignar a un conjunto de documentos de entrada un rótulo o etiqueta de acuerdo a categorías predefinidas y caracterizadas de alguna manera. La hipótesis de este trabajo se basa en la suposición de que aquel método de extracción que mejor resultado de clasificación obtenga conservará de mejor forma la esencia de las noticias. Intuitivamente, se puede ver que si el *gist* extraído se ajusta a la esencia del texto luego este será clasificado correctamente. En este sentido, se obtienen diferencias de rendimiento con las técnicas de extracción propuestas en la literatura.

El resto del artículo se encuentra dividido en las siguientes partes. A continuación, se presentan algunos trabajos relacionados al área de resumen automático y extracción de la idea principal de un texto. En el apartado 3 se describen los algoritmos evaluados. Luego, se expone el trabajo experimental, junto con algunas consideraciones surgidas. Finalmente, se concluye acerca del trabajo y se proponen algunas posibles líneas de trabajo.

2. Trabajos Relacionados

En la bibliografía existen una serie de trabajos de investigación que se basaron en el concepto de idea principal de un texto. Berger y Mittal [1] presentaron OCELOT, un sistema de extracción de la idea principal de páginas web, basándose en análisis de probabilidades. Los modelos que se utilizaron surgen de páginas web resumidas por humanos.

Tzoukermann [12] desarrolló una técnica denominada GIST-IT para resumir mensajes de correo electrónico. Ésta utiliza una combinación de métodos derivados de la lingüística y el aprendizaje automático a los efectos de extraer las frases más significativas de un texto.

GistSumm [9] es un algoritmo que permite realizar el resumen automático de textos. Utiliza una técnica que permite ponderar cada una de las oraciones de un texto en base a cuanto representa cada una la idea principal del texto. El método de resumen es extractivo y opera sobre documentos simples.

En cuanto a la utilización de resúmenes en la clasificación de textos, Shen y otros [10] propusieron un algoritmo de resumen aplicado específicamente a páginas web. Aprovecha la información de la estructura de los documentos y permite mejorar la precisión de la clasificación. Se obtienen mejoras respecto de los métodos tradicionales que operan sobre documentos de texto puro.

Kolcz y otros [3] propusieron la utilización de un proceso de resumen para la selección de características de un documento. En su trabajo, aplicaron técnicas de extracción junto a un conjunto de heurísticas para llegar al proceso final de clasificación.

3. Métodos de Extracción de la Idea Principal

Para la evaluación del proceso de extracción de la idea principal se utilizaron dos algoritmos clásicos del área y dos métodos basados en tomar oraciones particulares derivadas de la estructura de noticias cortas de prensa.

En particular, para el primer caso, se utilizaron los algoritmos de palabras clave y TF/ISF (Frecuencia del Término en la Oración / Frecuencia Inversa en el Texto). Para el segundo caso, se tomaron como oraciones *gist* de cada texto – por un lado – su título y – por el otro – la oración más "parecida" al título, de acuerdo a una métrica de distancia.

Método A: Palabras Clave (PC)

Este método se basa en el supuesto de que las mismas representan la idea principal de un texto [2]. Primero, se calcula la frecuencia de cada uno de los términos de un documento. Luego, por cada oración, se suma el peso individual de cada término (su frecuencia en el documento). El valor obtenido en la sumatoria representa el peso de cada oración. La oración de mayor peso representa la idea principal del texto.

Método B: TF/ISF

Frecuencia del Término en la Oración / Frecuencia Inversa en el Texto [4]. En este método se ponderan los términos de cada una de las oraciones. Para cada palabra W perteneciente a una oración O de un documento D su peso se calcula como:

$$p(W) = F(W,O) \bullet \log \frac{|O|_d}{DF(W)}$$

donde,

P(w)	= Peso de la palabra W.
F(W,O)	= Frecuencia del término W en la oración O.
$\|O\|_d$	= Cantidad de oraciones O en el documento d.
DF(W)	= Cantidad de oraciones en las cuales el término W ocurre en d.

Luego, se calcula el peso de cada oración promediando los pesos individuales de sus palabras.

$$p(O) = \frac{\sum_{i=1}^{|W|_O} p(W_i)}{|W|_O}$$

donde,

$|W|_O$ = Cantidad de términos W de la oración O.

Finalmente, de acuerdo a Larroca [4], aquella oración que obtenga mayor puntaje será la más representativa. Sobre este criterio, durante el trabajo experimental se encontró una diferencia sustancial, que sugiere que debe seleccionarse la oración de menor puntaje como aquella que representa el *gist*.

Método C: Oración correspondiente al Título del Texto (TITULO)

Esta elección se basa en la hipótesis de que el título de la noticia es altamente representativo del contenido y – por ende – esta oración es la que mejor representa la idea principal del texto. Esta suposición se fundamenta – además – en el hecho de que el título es definido por un humano capaz de interpretar el contenido. Sin embargo, los títulos – generalmente – son oraciones relativamente cortas, cuestión que se debe considerar.

Método D: Oración con mayor semejanza al Título del Texto (SIMTIT)

Es una variante con respecto a la técnica anterior, que intenta "salvar" la cuestión de la longitud del título. Manteniendo la hipótesis de que el título es un buen *gist* de un texto, el método consiste en seleccionar la oración de un texto que tenga mayor semejanza con éste. En estos experimentos se utilizó el coeficiente de Dice como métrica de semejanza, definido como:

$$DICE(O_1, O_2) \frac{2 \bullet \sum_{\forall Wi} F(w_{i,1}) \bullet F(w_{i,2})}{\sum_{\forall Wi} F(w_{i,1})^2 + \sum_{\forall Wi} F(w_{i,2})^2}$$

4. Trabajo Experimental

El objetivo del experimento consistió en evaluar distintos métodos de extracción de la oración que mejor representa la idea principal de un texto corto en español. Para ello, se definieron experimentos de campo basados en clasificación. Se utilizó una colección de noticias en español y un clasificador bayesiano clásico. Inicialmente, se evaluaron los métodos por separado y luego las combinaciones de los que resultaron más eficientes.

4.1. Datos de Prueba

El conjunto de prueba está formado por una colección de 1000 noticias monotemáticas en español, las cuales se encuentran divididas en 5 categorías temáticas: a) deportes, b) espectáculos, c) latinoamérica, d) política y economía argentina y e) política y economía internacional. Cada categoría contiene 200

noticias, donde cada una posee un título (la primera oración) y un conjunto de oraciones. El largo promedio de cada noticia es de 12 oraciones.

El preprocesamiento de la colección fue común para todos los métodos de extracción de la idea principal. Se normalizaron los caracteres llevándolos a minúsculas, luego se eliminaron las palabras vacías y finalmente se aplicó stemming, utilizando la herramienta para el español Snowball [11].

4.2. Clasificador

Debido a que el objetivo fue evaluar los algoritmos de selección de la idea principal mediante el proceso de clasificación, se seleccionó para los experimentos el clasificador popular naive Bayes, el cual ha demostrado alcanzar buena performance general [6]. Este clasificador se basa en la idea de utilizar las probabilidades conjuntas de los términos y las categorías para luego – aplicando la regla de Bayes – definir la categoría de un nuevo documento dado. En particular, se utilizó la implementación de la herramienta Bow [5] del clasificador naive Bayes, la cual fue entrenado con el 25% de las noticias completas de cada categoría, seleccionadas aleatoriamente en cada una de las corridas del test que se realizaron.

4.3. Métricas de Evaluación

La hipótesis de este trabajo se basa en la suposición de que aquél método de extracción que mejor resultado de clasificación obtenga conservará de mejor forma la esencia de las noticias. Esta es una evaluación extrínseca [7], es decir, se basa en cómo el método de extracción afecta a otra tarea, en este caso la clasificación.

Para la evaluación se utilizó la métrica normalizada Accuracy Percent (porcentaje de exactitud) de la herramienta Bow. Se eligió esta medida ya que no se está evaluando el algoritmo de clasificación sino el proceso, de acuerdo a diferentes datos de entrada que “intentan“ representar la misma noticia.

4.4. Resultados y Análisis

A los efectos de evaluar la performance de cada método de extracción de la idea principal se procedió a realizar las pruebas con cada uno, generando los extractos (*gists*), los cuales luego se utilizaron en la clasificación.

Para obtener una prueba de referencia, se realizó la clasificación de las noticias completas. Previamente, para validar la capacidad de la herramienta de clasificación, las noticias fueron categorizadas por un experto humano. El test con el software Bow arrojó un 94,13% de exactitud promedio en la clasificación, valor que se tomó como referencia (baseline) para los experimentos con los extractos de las noticias.

Además de evaluar los métodos de extracción de la idea principal descriptos en la sección 3, también se realizó la experiencia seleccionando una oración al azar. La inclusión de esta selección no se debió a que se la considera un método de extracción de *gist*, sino que permite apreciar más claramente la validez de los algoritmos utilizados.

En la tabla 1 se presentan los resultados de los experimentos descriptos. En la etapa de evaluación se encontraron diferencias de concepto con el método propuesto por Larocca [4] definido como TF/ISF. En tal trabajo, se plantea como *gist* del texto la selección de la oración cuyo valor de TF/ISF resulta máximo. Bajo esta apreciación y de acuerdo a las experiencias realizadas, el método resulta ineficiente (como se aprecia en los resultados). Es por ello, que se muestran los resultados del mismo método pero con la selección de la oración con valor máximo por un lado y aquella que arrojó el mínimo valor por el otro.

Método	**Exactitud (%)**	**Diferencia (%)**
Noticia completa (Baseline)	94,13	---
Palabras Clave (PC)	**86,40**	**8,21**
TF/ISF (Mínimo)	81,55	13,36
TF/ISF (Máximo)	64,56	31,41
TÍTULO	**86,80**	**7,79**
SIMTIT	**86,58**	**8,02**
Azar	66,24	29,63

Tabla 1 – Resultados de los diferentes métodos

Nótese que la mayor eficiencia la logran los métodos TÍTULO, SIMTIT y Palabras Clave, en este orden, que logran superar el 90% de eficiencia respecto del experimento de referencia. Se puede observar la baja eficiencia del método TF/ISF (máximo) tal cual se propone en [4], inclusive por debajo de la selección al azar. Debido a la naturaleza del método, el cual propone establecer una correspondencia con el método TF/IDF, ampliamente utilizado para ponderar los términos de un documento, surge una diferencia conceptual. En TF/IDF, el término IDF permite determinar la capacidad de discriminación de cada término respecto de los documentos de la colección. No obstante esta analogía no resulta homóloga en el caso de las oraciones de un texto debido a que se requiere establecer la relación de los términos de cada oración con el tema principal. En este sentido, se considera que existe un error de transcripción en el método propuesto en [4] y que la elección de la oración corresponde a la que obtiene el valor TF/ISF mínimo.

Dados estos resultados, se propuso realizar combinaciones de los métodos que mejores resultados arrojaron para tratar de mejorar la eficiencia. En la tabla 2 se presentan los resultados de la nueva clasificación utilizando dichas combinaciones.

Método	**Exactitud (%)**	**Diferencia (%)**
Noticia completa (Base)	94,13	---
SIMTIT + Palabras Clave	**91,01**	**3,31**
SIMTIT + TÍTULO	**90,93**	**3,40**
TITULO + Palabras Clave	**91,60**	**2,69**
TF/ISF (mínimo) + Palabras Clave	88,45	6,03
TF/ISF (mínimo) + TÍTULO	90,27	4,10
TF/ISF (mínimo) + SIMTIT	88,72	5,75

Tabla 2 – Resultados de los métodos combinados

Aquí se puede apreciar una mejora de aproximadamente del 5% entre cada uno de los mejores métodos, teniendo en cuenta que el método combinado incluye al mejor de los individuales. En este caso, hay que tener en cuenta que en esta segunda experiencia se tomaron 2 oraciones como la idea principal del texto.

5. Conclusiones y Trabajos Futuros

El tratamiento automático de información facilita a los usuarios la manipulación, evaluación y utilización de grandes cantidades de documentos. Una de las áreas involucradas es la que intenta identificar la información sustancial de un texto para generar una versión abreviada. Esta versión puede ser un resumen del texto, o bien, la oración que mejor representa la idea principal, denominada *gist*.

En este trabajo de investigación se evaluaron distintas técnicas de selección del *gist* de un texto corto monotemático escrito en español. Se utilizaron dos métodos clásicos de la literatura y dos propuestas extras. Para la evaluación se tomó el proceso de clasificación con al idea de determinar la capacidad del extracto para "representar" el contenido de la noticia y – por ende – su categoría temática.

Los resultados arrojaron que los métodos evaluados obtienen un buen comportamiento, superando – en algunos casos – el 90% de eficiencia respecto del experimento de referencia (noticia completa). Los mejores comportamientos se obtuvieron con TÍTULO y SIMTIT, seguidos por Palabras Clave.

Complementariamente, se realizaron experiencias utilizando combinaciones de los mejores métodos. En este caso, se logró un incremento de la eficiencia de un 5% utilizando conjuntamente los métodos TITULO y Palabras Clave. Esta situación es interesante si se acepta la utilización de 2 oraciones como idea principal.

Los resultados permiten establecer la validez de los métodos de *gist* presentados para el resumen de noticias en español. Su aplicación en el procesamiento de grandes volúmenes de información puede servir – por ejemplo – para presentar el contexto de un texto a los usuarios en un sistema de navegación de noticias.

A los efectos de enfocar la evaluación del proceso de extracción de la idea principal desde un punto de vista más real, se están planificando experimentos con jueces humanos, los cuales determinarán la capacidad de un algoritmo para obtener un extracto.

En este artículo, se han considerado textos cortos monotemáticos. No obstante, resulta una continuación interesante la validación y – eventualmente – modificación de alguno de estos algoritmos en textos más complejos como artículos científicos o reportes, donde previamente se requerirá de un análisis estructural más complejo que incluya la segmentación temática.

6. Referencias

[1] Berger, A. L. y Mittal, V.O. "*OCELOT:A system for summarizing web pages*". En: Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval. Athens, Greece, pp 144 – 151. 2000.

[2] Black, W.J. y Johnson, F.C. "*A Practical Evaluation of Two Rule-Based Automatic Abstraction Techniques*". Expert Systems for Information Management, Vol. 1, N. 3. Department of Computation. University of Manchester Institute of Science and Technology. 1998.

[3] Kolcz, A.; Prabakarmurthi, V. y Kalita, J. "*Summarization as Feature Selection for Text Categorization*". En: Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM-01). 2001.

[4] Larocca Neto, J., Santos, A.D., Kaestner, C.A.A., Freitas, A.A. *"Document clustering and text summarization"*. En: Proceedings of 4th International Conference Pratical Applications of Knowledge Discovery and Data Mining. 41–55. 2000.

[5] McCallum, A. "*Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering*". http://www.cs.cmu.edu/~mccallum/bow. 1996.

[6] McCallum, A. y Nigam, K. "*A comparison of event models for naive bayes text classification*". En: AAAI-98 Workshop on Learning for Text Categorization. 1998.

[7] Mani, I.; House, D.; Klein, G.; Hirschman, L.; Firmin, T. y Sundheim, B. "The TIPSTER SUMMAC Text Summarization Evaluation". En: Proceedings of EACL´99. Bergen, Noruega. 1999.

[8] Mani, I. y Maybury, M. (eds). Advances in Automatic Text Summarization. The MIT Press, Cambridge, Massachusetss. 1999.

[9] Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. "*GistSumm: A Summarization Tool Based on a New Extractive Method*". In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken, pp. 210-218 (Lecture Notes in Artificial Intelligence 2721). Springer-Verlag, Germany. 2003.

[10] Shen, D.; Chen, Z.; Yang, Q.; Zeng, H.; Zhang, B. Lu, Y. y Ma, W. "*Web-page Classification through Summarization*". En: Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval. Sheffield, United Kingdom, pp 242 – 249. 2004.

[11] Snowball. Spanish Stemming Algorithm.
http://www.snowball.tartarus.org/algorithms/spanish/stemmer.html

[12] Tzoukermann, M.E.; Muresan, S. y Klavans, J.L. "*GIST-IT: Summarizing Email Using Linguistic nowledge and Machine Learning*". En: Proceeding of the HLT and KM Workshop, EACL/ACL. 2001.