Unsupervised acquisition of a markov model for word correction using Wikipedia

*Rubén Dorado**

Fecha de recepción: 11 de agosto de 2014
Fecha de aprobación: 2 de octubre de 2014
Pp. 171-180

Abstract

This paper presents a work in progress on the area of automatic acquisition of corpora for spelling correction. Wikipedia contains a high quantity of information including relationships between concepts and named annotations. However, it also contains linguistic information such as misspellings written by many of the Wikipedia collaborators. In this paper, we propose an efficient method to analyze the link structure of Web-based dictionaries to construct a list of misspelled words and their corrections. The method is currently being researched and applied to the Wikipedia as a corpus.

Key words

Automatic acquisition, corpora, Wikipedia, Web-based dictionaries, corpus, misspellings.

* *Magister en Ciencias de la Computación, Universidad de Tokyo. Pregrado en Ingeniería de Sistemas, Universidad Nacional de Colombia..*

Adquisición no supervisada de un modelo Markov para corrección de texto utilizando Wikipedia

Resumen

Este artículo presenta un estudio en desarrollo acerca de la adquisición automática de corpora para realizar correcciones ortográficas. Wikipedia posee mucha información que incluye relaciones entre conceptos y anotaciones realizadas. Sin embargo, esta herramienta también contiene información lingüística acerca de errores ortográficos escritos por colaboradores en Wikipedia. En este artículo se propone un método eficiente para analizar el vínculo estructural de los diccionarios en línea para crear una lista de palabras mal escritas y sus correspondientes correcciones. Dicho método está bajo investigación y es utilizado en Wikipedia como un corpus.

Palabras clave

Adquisición automática, corpora, Wikipedia, diccionarios en línea, corpus, errores ortográficos.

Acquisition non supervisée d'un modèle de Markov pour correction Word à l'aide de Wikipedia

Résumé

Cet article présente un travail en cours dans le domaine de l'acquisition automatique de corpus pour la correction orthographique. Wikipedia contient un grand nombre d'informations y compris les relations entre concepts et annotations. Ce site fourni également des informations linguistiques telles que les fautes d'orthographe commises par les collaborateurs de Wikipedia. Dans cet article, nous proposons une méthode efficace pour l'analyse de la structure des dictionnaires en ligne et pour la construction de liste de mots mal orthographiés et leur correction. La méthode est actuellement à l'étude et appliquée comme corpus sur le site Wikipedia.

Mots-clés

Acquisition automatique, corpus, Wikipedia, dictionnaires en ligne, corpus, les fautes d'orthographe.

Aquisição não supervisada de um modelo de Markov para a correção ortográfica usando a Wikipédia

Resumo

Este trabalho apresenta um estudo em curso sobre a aquisição automática de corpora para correções ortográficas. Wikipédia tem muita informação, incluindo relações entre conceitos e anotações. No entanto, esta ferramenta também contém linguagem sobre erros ortográficos escritos por colaboradores de Wikipédia. Neste artigo se propõe um método eficiente para analisar a ligação estrutural dos dicionários on-line para criar uma lista de palavras com erros ortográficos e suas correções correspondentes. Este método está atualmente sob investigação e é usado na Wikipédia como um corpus.

Palavras-chave

Aquisição, corpora, Wikipédia, dicionários on-line, corpus, erros orto-gráficos.

1. Introducción

Nowadays there are an acceptable amount of linguistic resources for well-known languages such as English. However, there are still a lot of languages that suffer from the problem of insufficient resources to develop linguistic tools or research. The problem is even more dramatic since the creation of resources for new languages is not the only necessity, but to acquire linguistic information in a specific dialect or slang for specific areas is also required. For example, a tool may require a dictionary for a specific type of French that is spoken only in a particular country. One of the most important resources used in research and in practical tools such as web browsers and word processors is the spelling dictionaries.

Spelling errors are words in a text document that are mistakenly written when another one was intended. These kinds of errors are especially common in collective scenarios such as chat rooms, blogs and collaborative encyclopedias such as the well-known Wikipedia. Therefore, text taken from such resources is suitable to be used as a training data input in a system that can automatically detect misspellings as well as correct words. By detecting, it is implied that given the input the parameters of a statistical or mathematical model are discovered. Once the system has detected the problem, it should show a set of corrections.

The statistical model proposed for this work belongs to the markovian family. The Markov model is mixed with a probability distribution to calculate the probability of misspelled words given a history, which is explained in a further section of this paper. The objective of such method is to automatically acquire a set of misspelling candidates as well as a set of statistically correct words. The main motivation of training such model is that once the model has been trained, it can be used to create a misspelling dictionary.

The rest of this paper is organized as follows. Section 2 reviews works related to this study. Section 3 presents the statistical model proposed. Section 4 discusses the problems of using Wikipedia as a training data source. Section 5 presents the results achieved to date and the future work; and finally, section 6 brings the conclusions.

2. Previous works

Wikipedia has been previously used as training data to acquire different kind types of information in several works. To give some examples, Mihalcea (Mihalcea, 2007) describes a method to obtain semantic information for word sense disambiguation. Makris, Plegas, and Theodoridis (Makris, Plegas, & Theodoridis, 2013) presentspresent a system that trains himself itself with data taken from Wikipedia to perform text annotation and add semantic information to plan texts. Domínguez García, Rensing, and Steinmetz (Domínguez García, Rensing, & Steinmetz, 2011) extractsextract taxonomies from Wikipedia. Other works have probed proved that information such as hyperlink relations between a set of documents (Davison, 2000) can be used to obtain a set of interrelated terms. Other There are studies that have used hyperlinks in the Wikipedia since it provides more information than in aany other common web site (Nakayama, Hara, & Nishio, 2007) and more importantly, that such information can be used to create a thesaurus. However, Wikipedia articles also provide an article structure that can be taken into account to obtain semantic links between concepts.

For example, the article about John Von Neumann has the following sections: early life and education, career and abilities, personal life, later life and honors. In turn, the section career and abilities contains the following subsections: set theory, geometry, measure theory, ergodic theory, operator theory, lattice theory and game theory among others. In this case, it is possible to process the structure of the article and say that <set theory> is a <career and ability> of <John Von Newmann>. More than that, each text's subsection contains a set of keywords related to the specific term. The structure also defines a hierarchical model of related concepts.

On the other hand, automatic methods to acquire dictionaries have been studied for different languages such a Chinese (Chang, Lin, & Su, 1995) or Norwegian (Velupillai & Dalianis, 2008). The former usefirst one uses a probabilistic model mixed with elements of information theory such as perplexity to obtain a set of possible errors. It also proposes a method using linguistic information in the form of POS tags. The second work uses a method based on bilingual dictionaries and raw text corpora to add new words and a tool called GIZA++.

3. Proposed model

A statistical language model estimates thea probability distribution of for a set of words over all possible sentences. This set of words does not necessarily have to be in a sequence, it depends of on the model and the purpose of its useits intended purpose. Thus, a language model is a probability distribution over sentences:

$$w_1\ w_2\ w_3\ldots w_n:$$

$$p\ (w_1\ \ w_2\ w_3\ w_n\ ;\Theta\ ) \qquad (1)$$

where Θ represents additional information given to or required by the model such as part -of -speech, categories, syntax, context, semantic information, distributional parameters or hyper-parameters, etc.

N-gram models are the simplest method to represent language statistically. N-gram models use the n-1 preceding words to construct a probabilistic Markov model. This way, equation (1) can be transformed into a more convenient way using the probability theory to represent the joint probability as the combination of the conditional probability of words by making use of Bayes' rule:

$$p(w_1\ w_2\ w_3\ldots w_n\ )=p(w_1\ )p(w_2\ |w_1\ )\ p(w_3\ |w_1,w_2\ )\ldots\ p(w_n\ |w_1,w_2,\ldots,w_{n-1}) \qquad (2)$$

The main problem of this model is that although this representation is extremely simple, the number of conditional probabilities to be calculated is too high. The number of parameters required by the previous model with a corpus that contains a vocabulary of V=400 words. In order to calculate the probabilities for the bigram model, it is required to store V^2=16000 counts. It can be clearly seen that this is an exponential number is exponential and it is practically impossible to store such number of counts for full sentences.

The simplest approach to deal with this problem is to simplify this representation using a fixed value for n. For example, unigram models use n=1. This means unigram models do not take into account the history or preceding words. In such case, equation (2) becomes:

$$p(w_1 \, w_2 \, w_3 \ldots w_n) = p(w_1)p(w_2)p(w_3)\ldots\, p(w_n). \qquad (3)$$

Trigram models use n=3 and the model becomes:

$$p(w_1 \, w_2 \, w_3 \ldots w_n) = p(w_1)p(w_2 | w_1)p(w_3 | w_1 \, w_2)\ldots\, p(w_n | w_{n-2} \, w_{n-1}) \quad (4)$$

The acquisition of the probabilities for n-gram models is quite straightforward. There is, however, a problem when calculating the probability of a sentence. N-gram models suffer of sparseness of the data. In other words, many of the possible n-grams will not appear on the corpus since the combination of possibilities is too high. This is cumbersome because when calculating the probability of a particular sentence p(s), it is possible that an n-gram does not appear and in such case the probability of the n-gram is 0 and also so is the probability of the whole sentence. The objective is to train the parameters of such statistical models, or in other words, to calculate the probabilities and then to use them to obtain a set of words.

4. Results achieved and future work

The project is on the testing phase and experimenting with the probabilistic model. The projectIt started by building a tool that could read the content of the articles on Wikipediaof the Spanish version Wikipedia and then process each Wikipedia Article. The tool separates each Wikipedia article in a single file that can be processed in parallel. It also creates a list of article names for each file, and thus each article can be processed separately.

Additionally, a text preprocessing tool was also developed since text from Wikipedia articles contains additional information in the form of annotations. On this phase, HTML tags and Wikipedia annotations are removed to obtain the plain text, leaving only the desired information. Such text is used as an input to calculate n-gram probabilities as needed required by the model. At the moment, we are analyzing the best strategy to obtain the misspellings and corrections of the words with the probabilistic model.

4. References

Chang, J. -S., Lin, Y. -C., & Su, K. -Y. (1995). Automatic Construction of a Chinese Electronic Dictionary . Proceedings of the Third Workshop on Very Large Corpora, pp. 107-120.

Davison, B. D. (2000). Topical Locality in the Web. In Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval (SIGIR 2000). ACM Press.

Domínguez García, R., Rensing, C., & Steinmetz, R. (2011). Automatic acquisition of taxonomies in different languages from multiple Wikipedia versions. Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies (i-KNOW '11). New York: ACM.

Inkpen, D., & Islam, A. (2009). Real-Word Spelling Correction using Google Web 1T. Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009, pp. 1689-1692.

Makris, C., Plegas, Y., & Theodoridis, E. (2013). Improved text annotation with Wikipedia entities. Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC '13). New York: ACM.

Mihalcea, R. (2007). Using Wikipedia for Automatic Word Sense Disambiguation. Proceedings of NAACL HLT 2007. Rochester.

Nakayama, K., Hara, T., & Nishio, S. (2007). A Thesaurus Construction Method from Large Scale Web Dictionaries. 21st International Conference on Advanced Information Networking and Applications, 2007. AINA '07, pp. 932 - 939.

Velupillai, S., & Dalianis, H. (2008). Automatic Construction of Domain-specific Dictionaries on Sparse Parallel Corpora in the Nordic Languages. Coling 2008: Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization. Manchester: Association for Computational Linguistics ACL, pp. 10-16.