Information Processing and Management 000 (2016) 1–15

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

INFORMATION PROCESSING & MANAGEMENT
An International Journal

Assessing the impact of Stemming Accuracy on Information Retrieval – A multilingual perspective

Felipe N. Flores[*], Viviane P. Moreira

Instituto de Informática – UFRGS, Av. Bento Gonçalves, 9500, 91501-970 Porto Alegre, Brazil

ARTICLE INFO

Article history:
Received 31 March 2015
Revised 14 March 2016
Accepted 24 March 2016
Available online xxx

Keywords:
Stemming
Information Retrieval
Evaluation
Multilingual

ABSTRACT

The quality of stemming algorithms is typically measured in two different ways: (*i*) how accurately they map the variant forms of a word to the same stem; or (*ii*) how much improvement they bring to Information Retrieval systems. In this article, we evaluate various stemming algorithms, in four languages, in terms of accuracy and in terms of their aid to Information Retrieval. The aim is to assess whether the most accurate stemmers are also the ones that bring the biggest gain in Information Retrieval. Experiments in English, French, Portuguese, and Spanish show that this is not always the case, as stemmers with higher error rates yield better retrieval quality. As a byproduct, we also identified the most accurate stemmers and the best for Information Retrieval purposes.

1. Introduction

Stemming is the conflation of the variant forms of a word into a single representation, *i.e.*, the stem. For example, the terms *presentation, presenting*, and *presented* could all be stemmed to *present*. The stem does not have to be a valid word, but it needs to capture the meaning of the words.

Stemming is usually carried out by algorithms that strip word suffixes (but some also strip prefixes) which is why this technique is called affix stripping. Other stemming techniques include the use of dictionaries – which contain the correct form of stemming for the maximum number of words – and statistical stemming. Affix stripping stemmers are language dependent, that is, the rules are designed based on some knowledge of the language. One cannot use stemming rules designed for Portuguese, for example, and expect them to perform well on a French corpus. Statistical stemmers, however, aim at learning the stemming rules automatically and thus eliminating the need of knowing the language (Majumder et al., 2007; Paik, Mitra, Parui, & Järvelin, 2011).

The quality of stemming algorithms is typically assessed in one of these manners: (*i*) how correctly the stemmer maps semantically and morphologically related words to the same stem; or (*ii*) how much improvement the stemmer brings to Information Retrieval (IR). According to Jones and Galliers (1996) and Mollá and Hutchinson (2003), the first would be an *intrinsic evaluation* as it analyzes the accuracy of the results of the stemmer as a stand-alone system. The latter would be an *extrinsic evaluation*, because it analyses the impact of the stemmer in one of its applications.

Stemming is widely used in IR with the aim of increasing recall (*i.e.*, the number of relevant documents retrieved in response to a user query) (Baeza-Yates & Ribeiro-Neto, 2011). Another benefit is the reduction of the size of the index files, because a stem can represent many different words, resulting in fewer distinct index entries. Stemming is also used in other

[*] Corresponding author. Tel.: +55 5191186168.
E-mail address: fnflores@inf.ufrgs.br (F.N. Flores).

http://dx.doi.org/10.1016/j.ipm.2016.03.004

contexts, such as data mining, sentiment analysis, text categorization, automatic indexing, text summarization, information extraction, lexical analysis, and several natural language processing tasks. A number of studies report on the effectiveness of using stemming in an IR system, especially for the English language (Harman, 1991; Hull, 1996; Krovetz, 1993).

Since the goal of stemming is to increase recall, in practice, it tends to reduce precision as a side effect – an undesirable feature for Web search. Levene (2010) states that since stemming may not improve the top retrieved documents (*i.e.*, the ones that would appear on the first page of a search engine's result), it is not often used on the Web and large corpora. Still, the author mentions that major search engines such as Google and Yahoo! do employ some form of (light) stemming. Alternatively, rather than doing affix removal during indexing and querying, stemming can be done as a type of query expansion. Peng, Ahmed, Li, and Lu (2007) propose expanding the query with variant forms of the words and deal with stemming on a case by case basis, applying it selectively. The authors report gains in efficiency and quality of the retrieved results.

The interest in investigating stemming algorithms started back in 1960s and is still ongoing. While the first endeavors were devoted to create rule-based suffix strippers, nowadays the focus is on statistical stemmers, which demand no linguistic knowledge of the language for which they are designed. Also, as pointed out by Sharma (2012), once the suffixes to be removed are obtained, stemming can be done faster than applying rule-based stemmers.

Comparing stemming algorithms under different criteria is still performed in recent research. Jivani (2011) summarizes the advantages and disadvantages of known stemmers for English; while Moral, Antonio, Imbert, and Ramirez (2014) performed a more detailed analysis that also includes stemmers for other languages. Sirsat, Chavan, and Mahale (2013) evaluates the strength and accuracy of the most widely used English stemmers. Méndez-Cruz, Torres-Moreno, Medina-Urrea, and Sierra (2013) performed an extrinsic evaluation of stemmers on summarization tasks. The work by Brychcín and Konopík (2015) contemplates both intrinsic and extrinsic metrics in their experimental evaluation. The intrinsic evaluation relied on corpora annotated with the lemmas of the word forms and the extrinsic evaluations were on a standard IR setting. However, the goal was not to compare the two types of evaluation.

Although related, stemming and lemmatization are different tasks. While the former reduces words to their stems, the latter reduces them to their canonical forms (*i.e.*, dictionary form), or *lemmas*. For example, the word *having* would be lemmatized to *have* and stemmed to *hav*. The main difference in the two processes is that, while stemming can be done simply by applying a set of rules, lemmatization requires more complex tasks such as knowing the part-of-speech of the word and understanding its context in the sentence. Given its simplicity, stemming has been applied more widely than lemmatization.

In our earlier work Flores, Moreira, and Heuser (2010), we performed a comparison between the quality of a stemming algorithm and its effectiveness in an IR system for the Portuguese language. To the best of our knowledge, this was the first investigation on the relationship between these two quality indicators. Here we expand that study by adding English, French, and Spanish to the analysis and also by performing a topic-by-topic analysis, examining in how many topics which stemmers had a better result.

We experimented with various stemmers for English, French, Portuguese, and Spanish to measure their accuracy and also to assess the gain they bring to IR. Thus, as a byproduct, this paper identifies the most accurate stemmer for each language and the one that yields the biggest IR improvement.

An important aspect, as pointed by Paice (1994), is that looking at the values for extrinsic measures does not help the designer of the stemming algorithm in seeing where the mistakes are being made. In a recent survey, Moral et al. (2014) argues that extrinsic measures (such as precision and recall) are highly dependent on other tasks within the IR pipeline, and therefore they do not provide a good solution to evaluate the quality of the stemmers independently from other processes. Intrinsic measures, on the other hand, can pinpoint more clearly where the problems are and which improvements can be made. Moreover, it is important to study the two types of measurements together to understand how one impacts the other.

The remainder of this article is organized as follows: Section 2 discusses related work and introduces some background concepts for the evaluation of stemming algorithms; Section 3 presents the experiments that measure the accuracy of the stemmers; Section 4 describes the IR experiments done to compare the impact of the stemmers over retrieval effectiveness; Section 5 investigates the correlation between both quality indicators, and Section 6 concludes the article.

2. Background and related work

Paice (1996) proposed a method to evaluate the quality of stemmers using four intrinsic metrics:

- *Overstemming Index* (OI), which calculates the number of times a stemmer mistakenly removes part of the stem as if it were part of the suffix. This type of error will typically cause unrelated words to be combined, *e.g.* *news* and *new* are both stemmed to *new*. OI is zero when there are no overstemming errors and one when all words are stemmed to the same stem. In IR, a high OI will potentially lead to a decrease in precision, that is, many non-relevant documents would be retrieved by the query.
- *Understemming Index* (UI), which calculates the number of times a stemmer fails to remove a suffix. This type of error will typically prevent related words from being conflated, *e.g.* if *division* is stemmed to *divis* and *divide* is stemmed to *divid*. UI is zero when there are no understemming errors and one when no words are correctly combined by the stemmer. In

Please cite this article as: F.N. Flores, V.P. Moreira, Assessing the impact of Stemming Accuracy on Information Retrieval – A multilingual perspective, Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.03.004