bitext we help AI understand humans

Lemmatization versus stemming

1 Introduction

Working with large datasets is becoming more and more normal for companies no matter their size. The volume of the dataset can be a double edged-sword; on one hand, larger amount of data eases the possibility of finding the most accurate answer due the variety, but on the other hand the process of finding this answer becomes harder because the volume of data to review is wider.

Time is limited to find the appropriate answer so, to solve the problem maintaining the quality of the results we cannot reduce the size of the dataset. What we should do is enhance the search and indexing tools.

The idea of this paper is to explain how a lexical analyzer can improve the search outcomes by providing results that fit better with the query the user introduced.

2 Stemming and lemmatization for search

To address a solution for the problem proposed we should focus first in understanding how search engines work and stablish some parameters to measure the quality of the retrieved results.

In our normal use of language, words inflection is required. Depending on their function in the sentences, words will be written differently.

For the searching process these inflected words should be normalized, and this procedure can be done by lemmatization or stemming:

- **Stemming**: characters are removed of the end of the word by following language- specific rules.
- **Lemmatization**: based on its usage, the machine looks for the appropriate dictionary form of the word.

In weak inflected languages, such as English or Swedish using one or another methodology may not affect the quality of the results, however in most European languages the percentage of error will be increased if the search is conducted using stemming.

On the other hand, the quality of a search depends on two things:

- *Recall* is the fraction of results that are relevant or correct.
- *Precision or sensitivity* is the amount of correct results that are retrieved.

The ideal should be choosing the method among the two explained that doesn't reduce any of the previous parameters.

3 How lemmatization works

In inflected languages search faces sometimes the problem of ambiguous words. The same word can have different meanings depending on the context. And here is where lemmatization is key.

To determine the lemma of the word it takes into consideration its intended meaning, depending on the context the tool will automatically determine which is the right lemma for the word and therefore the results retrieved for a query will be more accurate.

If we focus on the stem it can happen that an ambiguous word with two different meaning will have the same stem. In those cases, stemming will not help with ambiguity and the returned results will include irrelevant noise.

4 Examples

Let's see some examples in different languages so we can see the differences of using stemming and lemmatization:

4.1 French

Input	Stem	Lemma
Livre *verb: to deliver*	livr-	livrer
Livre *noun: book*	livr-	livre
Somme *verb: to add up*	somm-	sommer
Somme *noun: total*	somm-	somme
Coté *verb: to quote*	cot-	coter
Coté *adj: popular*	cot-	coté

4.2 Spanish

Input	Stem	Lemma
Arma *verb: to arm*	arm	armar
Arma *noun: weapon*	arm	arma
Pienso *verb: to think*	piens	pensar
Pienso *noun: fodder*	piens	pienso
Traje *verb: to buy*	traj	traer
Traje *noun: dress*	traj	traje

4.3 Polish

Input	Stem	Lemma
Gra *verb: to play*	-gra-	grać
Gra *noun: play*	-gra-	gra
Bez *prep: without*	bez	bez
Bez *noun: elder*	bez-	bez
Koło *prep: next to*	koło	koło
Koło *noun: wheel*	koł-	koło

4.4 Russian

Input	Stem	Lemma
для *prep: for*	для	длить
для *verb: perpetuate*	-дл-	для
знать *verb: to know*	-зна-	знать
знать *noun: aristocracy*	-знать-	знать
стать *verb: become*	-ст-	стать
стать *noun: physical shape*	-стать-	стать

4.5 German

Input	Stem	Lemma
Macht *noun: power*	match	Macht
Macht *verb: to do*	match	machten
Lauf *noun: ride*	lauf	Lauf
Lauf *verb: to run*	lauf	Laufen
Sang *noun: song*	sang	Sang
Sang *verb: to sing*	sang	singen

5 Conclusion

As we can observe, an ambiguous pair of words in most of the inflected languages will share the same stem but will have a different lemma.

In those cases, the only way of getting accurate results is by use of lemmatization. This is possible for the software because it takes into account the context of the word and therefore it achieves real understanding of its meaning.