Midterm project report for the Natural Language Understanding Systems course

Anonymous ACL submission

Abstract

This work provides a concept tagging tool for queries related to movies, such as ”who is the director of thor”. The final version of the script enables tweaking of various learning parameters to better evaluate which setting is the best for the situation.

1 Introduction

2 Data set analysis

The dataset is subdivided into a training set and a test set. These sets are in a word per line format, with the sentences being separated by an empty line. Multiple columns represent the features of the specific word (such as part-of-speech tag and the lemma) and the correct concept tag.

An analysis of the distribution of the data provided was deemed necessary to ensure it’s correct usage.

2.1 Zipf’s law

The first analysis performed on the training data is to check whether or not Zipf’s law is verified. Due to the nature of the data, a bias towards words typical of the movie industry was expected. This can be seen clearly in the word frequency distribution, with ”movies” and ”movie” occupying respectively the second and fifth spot. (with ”movies” appearing more frequently than ”of”)

2.2 Concept distribution

The most frequent concept is ”movie.name” by a wide margin, which means this plus out-of-span tags make up 86% of the tags present in the training dataset.

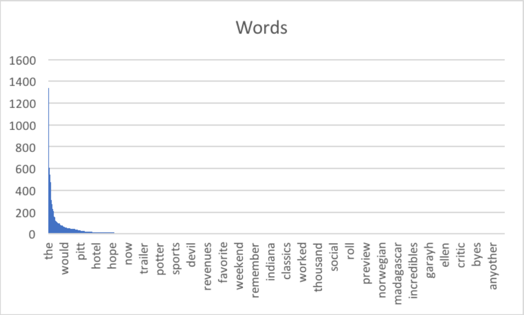

Figure 1: Zipf’s law

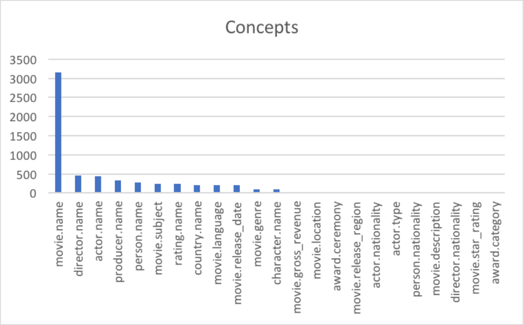

Figure 2: Concept distribution

3 Baseline solution

The case of a single transducer from word to concept will be taken in consideration as the baseline for all the proposed solutions. This doesn’t take into consideration the structure of the phrase so it is expected to be very inaccurate.

```
accuracy:  67.32%
precision:  20.15%
recall:  54.72%
FB1:  29.45
```

The resulting F1-Score is very low as expected.

4 Direct solution

The simplest and most direct method is obtained by simply composing the transducer from the previous step with the concept model trained on the training data. To do this the training data was converted from word-per-row to a concept sentenceper-row format. These sentences have been used to train a wide array of language models, by changing the order of the n-grams and the smoothing method.

It can be observed that using a language model trained on bigrams achieves better performance than one trained on trigrams, which can be explained by looking at the training and test sets.

4.1 Witten bell, 3-grams

```
accuracy:  92.62%
precision:  76.58%
recall:  74.61%
FB1:  75.58
```

4.2 Witten bell, 2-grams

```
Smoothing method: witten_bell
Order: 2-grams
accuracy:  92.68%
precision:  78.51%
recall:  74.34%
FB1:  76.37
```

5 Solution using extra features

The next experiment was to try and include the extra features included in the dataset to improve the F1-Score. At a first glance the chaining of multiple transducers, each passing from one feature to the next, seemed like a good idea. Because of this 3 transducers were built and composed: word-to-lemma, lemma-to-part of speech, part of speech-to-concept tag. By composing the result with the concept model it is shown to be even worse than the baseline:

```
Smoothing method: witten_bell
Order: 2-grams
accuracy:  72.81%
precision:  27.35%
recall:   6.14%
FB1:  10.03
```

An explanation can be found by considering that this method is reducing the input space of the final transducer from the size of the vocabulary to just 38 pos-tags.

Because of this, a second method which keeps the first transducer, while replacing the second and third with a lemma-to-concept transducer. The second transducer takes into consideration the pos-tags in the calculation of the weights.

```
Smoothing method: witten_bell
Order: 3-grams
accuracy:  92.34%
precision:  76.06%
recall:  73.97%
FB1:  75.00
```

The performance of this method is comparable to the version without extra features.

6 Solution with generalization

7 Solution with generalization and extra features

8 Using frequency cut-off

9 Conclusion

9.1 Sections

If you are using the provided LaTeX and BibTeX style files, you can use the command `\citet` (cite in text) to get "author (year)" citations.

If the BibTeX file contains DOI fields, the paper title in the references section will appear as a hyperlink to the DOI, using the hyperref LaTeX package. To disable the hyperref package, load the style file with the `nohyperref` option: `\usepackage[nohyperref]{acl2017}`

As examples, we cite (Goodman et al., 2016) to show you how papers with a DOI will appear in the bibliography. We cite (Harper, 2014) to show how papers without a DOI but with an ACL Anthology Identifier will appear in the bibliography.

As reviewing will be double-blind, the submitted version of the papers should not include the authors' names and affiliations. Furthermore, self-references that reveal the author's identity, *e.g.*,

> "We previously showed (Gusfield, 1997) ..."

should be avoided. Instead, use citations such as

> "Gusfield (1997) previously showed ... "

Please do not use anonymous citations and do not include acknowledgements when submitting your papers. Papers that do not conform to these requirements may be rejected without review.

References: Gather the full set of references together under the heading **References**; place the section before any Appendices, unless they contain references. Arrange the references alphabetically by first author, rather than by order of occurrence in the text. Provide as complete a citation as possible, using a consistent format, such as the one for *Computational Linguistics* or the one in the *Publication Manual of the American Psychological Association* (American Psychological Association, 1983). Use of full names for authors rather than initials is preferred. A list of abbreviations for common computer science journals can be found in the ACM *Computing Reviews* (for Computing Machinery, 1983).

The LaTeX and BibTeX style files provided roughly fit the American Psychological Association format, allowing regular citations, short citations and multiple citations as described above.

Appendices: Appendices, if any, directly follow the text and the references (but see above). Letter them in sequence and provide an informative title: **Appendix A. Title of Appendix**.

References

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. *Computing Reviews* 24(11):503–512.

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1–11. https://doi.org/10.18653/v1/P16-1001.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Mary Harper. 2014. Learning from 26 languages: Program management and science in the babel program. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, page 1. http://aclweb.org/anthology/C14-1001.

A Supplemental Material

ACL 2017 also encourages the submission of supplementary material to report preprocessing decisions, model parameters, and other details necessary for the replication of the experiments reported in the paper. Seemingly small preprocessing decisions can sometimes make a large difference in performance, so it is crucial to record such decisions to precisely characterize state-of-the-art methods.

Nonetheless, supplementary material should be supplementary (rather than central) to the paper. **Submissions that misuse the supplementary material may be rejected without review.** Essentially, supplementary material may include explanations or details of proofs or derivations that do not fit into the paper, lists of features or feature templates, sample inputs and outputs for a system, pseudo-code or source code, and data. (Source code and data should be separate uploads, rather than part of the paper).

The paper should not rely on the supplementary material: while the paper may refer to and cite the supplementary material and the supplementary material will be available to the reviewers, they will not be asked to review the supplementary material.

Appendices (*i.e.* supplementary material in the form of proofs, tables, or pseudo-code) should come after the references, as shown here. Use `\appendix` before any appendix section to switch the section numbering over to letters.

B Multiple Appendices

. . . can be gotten by using more than one section. We hope you won’t need that.