Language Understanding System: Mid-term Project

Montagner Andrea
andrea.montagner@studenti.unitn.it

Abstract

The aim of this project is to develop o Spoken Language Understanding module for the Movie Domain capable of performing PoS tagging. In order to achieve this goal a language model is firstly created and then trained to predict unseen utterances from a testing pool of data. In the following sections multiple approaches with respective solutions are presented, each one evaluated in term of accuracy, precision, recall and f1 score. Finally these results are compared to each other in order to determine the best approach ti this problem.

1 Introduction

A Concept Tagger is a module which is capable of assigning a concept, in this case a tag, to words belonging to a sentence. This operation, though, requires the presence of a dataset composed by couples (words, tag) large enough in order to have a sufficient number of utterances to perform both the training and testing phase.
In particular, the material that was provided for this project was:

- A train dataset, both for the Movie Domain and for the additional features,
- A test dataset, both for the Movie Domain and for the additional features,
- A Perl script, `"conlleval.pl"`, to evaluate the performances in terms of accuracy, precision, recall and f1 score.

2 Data Analysis

The provided dataset is the Microsoft NL-SPARQL dataset and it was already split into two parts: one for training and one for testing. In addition each of these groups of data was divided into main data for the Movie Domain and additional features.
To summarize, what was given were four files:

- **NL-SPARQL.train.data** containing a two columns set of data, tab separated, where the first column represent the words and the second the tags,
- **NL-SPARQL.train.feats.txt** containing a three columns set of data, tab separated, where the first column represent the words and the second the tags and third the lemmas,
- **NL-SPARQL.test.data** containing a two columns set of data, tab separated, where the first column represent the words and the second the tags,
- **NL-SPARQL.test.feats.txt** containing a three columns set of data, tab separated, where the first column represent the words and the second the tags and third the lemmas.

In these files, each line contains a word (with the respective tag/lemma) of a sentence and at the end of each sentence there is an empty line as separator from the next.
Some examples are shown below:

DATA EXAMPLE

Analysis of the distributions, as far as the provided files are concerned, can be performed grouping for words as well as for tags. In the two following subsections, both analysis are shown.

2.1 IOB Definition & Distribution

Concept Tags are here represented using an IOB notation, which is a common tagging for-

mat when dealing with ...
The used prefixes are:

- **B** - Beginning of a span
- **I** - Inside of a span
- **E** - End of a span
- **O** - Outside of a span

As long as IOBs distributions is concerned, considering that the dataset is Movie Domain, the analysis mirrors the prediction and outputs as the most frequent tag `movie.domain` divided in `I.movie.domain` and `B.movie.domain`.

In this analysis the tag `O` is not taken into consideration because it can be considered a "stop-word" for tags.
TABLE

2.2 Words Distribution

Very similar are results for the analysis on the word frequency. Almost all of most frequent words are identified as stop words, like `the`, `of`, `in`. Nonetheless there are exceptions, like `movies` and `movie`, which are relevant to understand the topic of the considered dataset.

2.3 Additional Features

2.4 Layout

Format manuscripts two columns to a page, in the manner these instructions are formatted. The exact dimensions for a page on A4 paper are:

- Left and right margins: 2.5 cm
- Top margin: 2.5 cm
- Bottom margin: 2.5 cm
- Column width: 7.7 cm
- Column height: 24.7 cm
- Gap between columns: 0.6 cm

Papers should not be submitted on any other paper size. If you cannot meet the above requirements about the production of your electronic submission, please contact the publication chairs above as soon as possible.

Type of Text	Font Size	Style
paper title	15 pt	bold
author names	12 pt	bold
author affiliation	12 pt	
the word "Abstract"	12 pt	bold
section titles	12 pt	bold
subsection titles	11 pt	bold
document text	11 pt	
captions	11 pt	
abstract text	11 pt	
bibliography	10 pt	
footnotes	9 pt	

Table 1: Font guide.

2.5 Fonts

For reasons of uniformity, Adobe's **Times Roman** font should be used. In LaTeX2e this is accomplished by putting

```
\usepackage{times}
\usepackage{latexsym}
```

in the preamble. If Times Roman is unavailable, use **Computer Modern Roman** (LaTeX2e's default). Note that the latter is about 10% less dense than Adobe's Times Roman font.

2.6 The First Page

Center the title, author name(s), and affiliation(s) across both columns (or, for the initial submission, **Anonymous ACL submission** for names and affiliations). Do not use footnotes for affiliations. Include the paper ID number assigned during the submission process in the header. Use the two-column format only when you begin the abstract.

Title: Place the title centered at the top of the first page, in a 15-point bold font. (For a complete guide to font sizes and styles, see Table 1) Long titles should be typed on two lines without a blank line intervening. Approximately, put the title at 2.5 cm from the top of the page, followed by a blank line, then the author name(s), and the affiliation(s) on the following line. Do not use only initials for given names (middle initials are allowed). Do not format surnames in all capitals (*e.g.*, use "Mitchell" not "MITCHELL"). Do not format title and section headings in

Command	Output
{\"a}	ä
{\^e}	ê
{\`i}	ì
{\.I}	İ
{\o}	ø
{\'u}	ú
{\aa}	å

Command	Output
{\c c}	ç
{\u g}	ğ
{\l}	ł
{\~n}	ñ
{\H o}	ő
{\v r}	ř
{\ss}	ß

Table 2: Example commands for accented characters, to be used in, *e.g.*, BibTeX names.

all capitals as well except for proper names (such as "BLEU") that are conventionally in all capitals. The affiliation should contain the author's complete address, and if possible, an electronic mail address. Start the body of the first page 7.5 cm from the top of the page.

The title, author names and addresses should be completely identical to those entered to the electronic paper submission website in order to maintain the consistency of author information among all publications of the conference. If they are different, the publication chairs may resolve the difference without consulting with you; so it is in your own interest to double-check that the information is consistent.

Abstract: Type the abstract at the beginning of the first column. The width of the abstract text should be smaller than the width of the columns for the text in the body of the paper by about 0.6 cm on each side. Center the word **Abstract** above the body of the abstract using the font size and style shown in Table 1. The abstract should be a concise summary of the general thesis and conclusions of the paper. It should be no longer than 200 words. The font size of the abstract text should be as shown in Table 1.

Text: Begin typing the main body of the text immediately after the abstract, observing the two-column format as shown in the present document. Do not include page numbers in the final version.

Indent: Indent when starting a new paragraph, about 0.4 cm.

2.7 Sections

Headings: Type and label section and subsection headings in the style shown on the present document. Use numbered sections (Arabic numerals) in order to facilitate cross references. Number subsections with the section number and the subsection number separated by a dot, in Arabic numerals. Do not number subsubsections (*i.e.*, use `\subsubsection*` instead of `\subsubsection`).

Citations: Citations within the text appear in parentheses as (Gusfield, 1997) or, if the author's name appears in the text itself, as Gusfield (1997). Using the provided LaTeX style, the former is accomplished using `\cite` and the latter with `\shortcite` or `\newcite`. Collapse multiple citations as in (Gusfield, 1997; Aho and Ullman, 1972); this is accomplished with the provided style using commas within the `\cite` command, *e.g.*, `\cite{Gusfield:97,Aho:72}`. Append lowercase letters to the year in cases of ambiguities. Treat double authors as in (Aho and Ullman, 1972), but write as in (Chandra et al., 1981) when more than two authors are involved.

Also refrain from using full citations as sentence constituents. We suggest that instead of

> "(Gusfield, 1997) showed that ..."

you use

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

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.

You can use the command `\citealp` (alternative cite without parentheses) to get "author year" citations (which is useful for using citations within parentheses, as in Gusfield, 1997).

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]{acl2018}`.

Compilation Issues: Some of you might encounter the following error during compilation:

output	natbib	previous ACL style files
(Gusfield, 1997)	\citep	\cite
Gusfield (1997)	\citet	\newcite
(1997)	\citeyearpar	\shortcite

Table 3: Citation commands supported by the style file. The citation style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.

"\pdfendlink *ended up in different nesting level than* \pdfstartlink."

This happens when pdflatex is used and a citation splits across a page boundary. To fix this, disable the hyperref package (see above), recompile and see the problematic citation. Next rewrite that sentence containing the citation. (See, *e.g.*, http://tug.org/errors.html)

Digital Object Identifiers: As part of our work to make ACL materials more widely used and cited outside of our discipline, ACL has registered as a CrossRef member, as a registrant of Digital Object Identifiers (DOIs), the standard for registering permanent URNs for referencing scholarly materials. We are requiring all camera-ready references to contain the appropriate DOIs (or as a second resort, the hyperlinked ACL Anthology Identifier) to all cited works. Thus, please ensure that you use BibTeX records that contain DOI or URLs for any of the ACL materials that you reference. Appropriate records should be found for most materials in the current ACL Anthology at http://aclanthology.info/.

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 acknowledgments 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* (Association 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**.

2.8 Footnotes

Footnotes: Put footnotes at the bottom of the page and use the footnote font size shown in Table 1. They may be numbered or referred to by asterisks or other symbols.[1] Footnotes

[1] This is how a footnote should appear.

should be separated from the text by a line.[2]

2.9 Figures and Tables

Placement: Place figures and tables in the paper near where they are first discussed, as close as possible to the top of their respective column.

Captions: Provide a caption for every illustration; number each one sequentially in the form: "Figure 1: Caption of the Figure." "Table 1: Caption of the Table." Type the captions of the figures and tables below the body, using the caption font size shown in Table 1.

2.10 Equation

An example equation is shown below:

$$A = \pi r^2 \tag{1}$$

The numbering (if any) and alignment of the equations will be done automatically (using `align` or `equation`).

2.11 Accessibility

In an effort to accommodate the color-blind (as well as those printing to paper), grayscale readability for all accepted papers will be encouraged. Color is not forbidden, but authors should ensure that tables and figures do not rely solely on color to convey critical distinctions. A simple criterion: All curves and points in your figures should be clearly distinguishable without color.

3 Evaluation: Baseline method

It is also advised to supplement non-English characters and terms with appropriate transliterations and/or translations since not all readers understand all such characters and terms. Inline transliteration or translation can be represented in the order of: original-form transliteration "translation".

4 Other approaches: different training methods

The ACL 2018 main conference accepts submissions of long papers and short papers. Long papers may consist of up to eight (8) pages of content plus unlimited pages for references. Upon acceptance, final versions of long papers will be given one additional page – up to nine (9) pages of content plus unlimited pages for references – so that reviewers' comments can be taken into account. Short papers may consist of up to four (4) pages of content, plus unlimited pages for references. Upon acceptance, short papers will be given five (5) pages in the proceedings and unlimited pages for references.

For both long and short papers, all illustrations and tables that are part of the main text must be accommodated within these page limits, observing the formatting instructions given in the present document. Supplementary material in the form of appendices does not count towards the page limit.

However, note that supplementary material should be supplementary (rather than central) to the paper, and that reviewers may ignore supplementary material when reviewing the paper (see Appendix A). Papers that do not conform to the specified length and formatting requirements are subject to be rejected without review.

Workshop chairs may have different rules for allowed length and whether supplemental material is welcome. As always, the respective call for papers is the authoritative source.

5 Error Analysis

6 Discussion

Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowledgments section (*i.e.*, use `\section*` instead of `\section`). Do not include this section when submitting your paper for review.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

[2]Note the line separating the footnotes from the text.

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

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

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

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)*, pages 1–11. Association for Computational Linguistics.

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*, page 1. Dublin City University and Association for Computational Linguistics.

A Supplemental Material

ACL 2018 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.