Project report

Machine Learning Project

Antonio Toral
Information Science

week 6

Project report

- Conduct a small-scale research on a topic related to the course.
- Write a report:
  - Formulation and motivation of the research question.
  - Discussion of the relevant scientific literature (at least 3 articles).
  - Technical aspects of the performed experiment and / or implementation.
  - Report and discussion of the results.
  - Length: max. 10 pages, excluding any attachments.
- Presentation in week 9.
- In groups of 2 or 3 (same group as for the project proposal)

Topics

- Ideally same as in the research proposal
- or another (still related to the course!) ...
- In any case: state clearly how the project relates to the research proposal.
- Implementation-wise: more than your code for the assignment on that topic. In other words, the same functionality just applying it to another dataset is not enough.

Word processing

LaTeX

Word processing

- LaTeX is a document markup language.
- For a good tutorial, check: `http://www.latex-tutorial.com/tutorials/`
- Suitable editors include Overleaf (on-line), Texmaker and TeXworks (Linux, OS X and Windows) and Gummi (Linux and Windows).
- Gummi and Overleaf have a live preview: the pdf is displayed without needing to compile manually. As you change the LaTeXcode, the preview is automatically updated.

Word processing

Other options (but LaTeX is recommended)

LibreOffice
The Document Foundation

OpenOffice.org
The Open Source Office Suite

Office 365

Using references

Recently, machine learning approaches have been explored to estimate the age of an author or speaker using text uttered or written by the person. This has been modeled as a classification problem, ... In machine learning research, these cohorts have typically been determined for practical reasons relating to distribution of age groups within a corpus, ... For example, researchers have modeled age as a two-class classification problem with boundaries at age 40 (Garera and Yarowsky, 2009) or 30 (Rao et al., 2010). Another line of work has looked at modeling age estimation as a three-class classification problem (Schler et al., 2006; Goswami et al., 2009)... As an example of one of these studies, Pennebaker and Stone (2003) analyzed the relationship between language use and aging by collecting data from a large number of previous studies. They used LIWC (Pennebaker et al., 2001) for analysis. They found that with increasing age, people tend to use more positive and fewer negative affect words, more future-tense and less past-tense, and fewer self-references.... Age classification experiments have been conducted on a wide range of types of data, including ... Twitter (Rao et al., 2010). Effective features were both content features (such as unigrams, bigrams and word classes) as well as stylistic features (such as part-of-speech, slang words and average sentence length). These separate published studies present some commonalities of findings. However, based on these results from experiments conducted on very different datasets, it is not possible to determine how generalizable the models are. Thus, there is a need for an investigation of generalizability specifically in the modeling of linguistic variation related to age, which we present in this paper.

Using references

As corpus, we used an 80 million word newspaper Dutch corpus, a subset of the Twente Newspaper corpus (Ordelman et al., 2007).[1] For comparison with spoken language, we used the Corpus of Spoken Dutch (Oostdijk, 2000).

This is in strong contrast with similar constructions in English, i.e. the optional presence of *that* in finite complement clauses and relatives, which has been the subject of numerous studies (see, among others, Ferreira and Dell (2000), Hawkins (2002), and Wasow et al. (2011)).

In particular, Roland et al. (2006) observe that the strongest predictor for complementizer presence is the governing verb. Jaeger (2010) extends this result by showing that this effect can to a large extent (but not completely) be contributed to subcategorization frequency, i.e. the likelihood that a governing verb occurs with or without a complement clause.

[1] material is from *Algemeen Dagblad* and *NRC Handelsblad*, 1994 and 1995

Using references

Two types

- If the reference is not part of the sentence, it is written fully between parentheses:
  . . . a subset of the Twente Newspaper corpus (Ordelman et al., 2007).
- If the reference is part of the sentence, the author name(s) is written as part of the sentence, and the year between parentheses:
  In particular, Roland et al. (2006) observe that ....

LaTeX and BibTex

- Tracking references and lists of literature is time consuming.
- There are several packages to assist you
- LaTeX uses BibTex:
    - Create a database with all your references
    - In the text you use the key of a paper/book
    - At the end of your text LaTeX will automatically include a list of the references that you have cited

Bibliography in Word, LibreOffice, OpenOffice

Note: support only provided for BibTex

- For Microsoft Word see: Reference Manager and RefWorks
- For Microsoft Word see also:

  http://www.scribendi.com/advice/how_to_create_a_bibliography_using_word.en.html
- For LibreOffice see:

  https://help.libreoffice.org/Writer/Creating_a_Bibliography/nl

  https://help.libreoffice.org/Writer/Insert_Bibliography_Entry

  If you want to use BibTex lists in LibreOffice via JabRef see:

  http://onetransistor.blogspot.nl/2015/04/libreoffice-bibliography-jabref.html
- For OpenOffice see:

  https://wiki.openoffice.org/wiki/Documentation/OOoAuthors_User_Manual/Writer_Guide/Creating_a_bibliography

Using references

```
As corpus, we used an 80 million word newspaper Dutch corpus,
a subset of the Twente Newspaper corpus \citep{ordelman:2007}.

In particular, \citet{roland:2006} observe that the strongest
predictor for complementizer presence is the governing verb.
```

bibtex entry

```
@article{Roland:2006,
   author        = "Roland, D. and J.L. Elman and V.S. Ferreira",
   title         = "Why is that? {S}tructural prediction and ambiguity
           resolution in a very large corpus of English sentences",
   journal       = "Cognition",
   volume        = "98",
   number        = "3",
   pages         = "245--272",
   year          = "2006"
}
```

bibtex entry

```
@inproceedings{Wasow:2011,
   author         = "Wasow, T. and T.F. Jaeger, and D.M. Orr",
   title          = "Lexical variation in relativizer frequency",
   booktitle      = "Proceedings of the workshop on Expecting
          the Unexpected: Exceptions in Grammar",
   editors        = "H. Wiese and H. Simon",
   publisher      = "Mouton De Gruyter",
   pages          = "",
   year           = "2011",
}
```

bibtex entry

```
@book{Baayen:2008,
   author    = "Baayen, R.H.",
   title     = "Analyzing Linguistic Data",
   year      = "2008",
   publisher = "Cambridge University Press"
}
```

Google Scholar and bibtex

- Your bibtex file is a plain text file with extension .bib, so for example references.bib
- You can use any text editor to edit a bibtex file
  - Gummi and Texmaker both support the bibtex format
  - Special editors for bibtex:
    - kbibtex (Linux)
    - JabRef (Linux, OS X, Windows)
- We will see more about the formatting of bibtex entries in the following examples. To know more you can check the documentation.
- Google Scholar can produce bibtex entries for you
  - Under an article click on the link “cite”. Then choose “BibTeX”.

Working with bibtex

```
\usepackage[round]{natbib}

\begin{document}

 ...\citep{} ...\citet{}...

\bibliographystyle{plainnat}
\bibliography{my_bibtex_file}

\end{document}
```

Working with bibtex

- Place my_bibtex_file.bib in the same folder as your main LaTeX text file (text file with extension `.tex`).
- Create the PDF for you text (this also creates a file with references) as follows:
  1. Compile the PDF.
  2. Run the command bibtex.
  3. Compile the PDF 2x again: now in the PDF you should see all the references and a bibliography.
  4. Or use `Overleaf`. It compiles everything for you at once.

Details

More citing options:

`\citet{jon90}`	Jones et al. (1990)
`\citet[chap. 2]{jon90}`	Jones et al. (1990, chap. 2)
`\citep{jon90}`	(Jones et al., 1990)
`\citep[chap. 2]{jon90}`	(Jones et al., 1990, chap. 2)
`\citep[see][]{jon90}`	(see Jones et al., 1990)
`\citep[see][chap. 2]{jon90}`	(see Jones et al., 1990, chap. 2)
`\citet*{jon90}`	Jones, Baker, and Williams (1990)
`\citep*{jon90}`	(Jones, Baker, and Williams, 1990)

For even more citing options see:

`http://merkel.zoneo.net/Latex/natbib.php`

References

V.S. Ferreira and G.S. Dell. Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, 40(4):296–340, 2000.

J.A. Hawkins. Symmetries and asymmetries: their grammar, typology and parsing. *Theoretical Linguistics*, 28(2):95–150, 2002.

Florian Jaeger. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61:23–62, 2010.

Nelleke Oostdijk. The spoken dutch corpus: Overview and first evaluation. In *Proceedings of LREC 2000*, pages 887–894, 2000.

Roeland Ordelman, Franciska de Jong, Arjan van Hessen, and Hendri Hondorp. Twnc: a multifaceted dutch news corpus. *ELRA Newsletter*, 12(3/4):4–7, 2007.

D. Roland, J.L. Elman, and V.S. Ferreira. Why is that? Structural prediction and ambiguity resolution in a very large corpus of english sentences. *Cognition*, 98(3): 245–272, 2006.

T. Wasow, T.F. Jaeger, and D.M. Orr. Lexical variation in relativizer frequency. In *Proceedings of the workshop on Expecting the Unexpected: Exceptions in Grammar*. Mouton De Gruyter, 2011.