Statistical Machine Translation
Referat

Alexander Fraser
CIS, LMU München

2016.11.15 SMT and NMT

Quick Question

- How many of you will take the "Erweiterungsmodul" next semester (typically 2nd semester of Masters)?

Schein in this course

- Referat (next slides)
- Hausarbeit
  - 6 to 10 pages (an essay/prose version of the material in the slides), due 3 weeks after the Referat

Referat Topics

- We should have about 1-4 literature review topics and 6-9 projects
  - Projects will hold a Referat which is a mix of literature review/motivation and own work

Referat Topics - II

- Literature Review topics
  - (S)MT with little parallel data (underresourced languages)
  - Language modeling (compare count-based with smoothing and neural language models)
  - Tuning log-linear models: MIRA
  - Basic word-sense disambiguation and WSD approaches to SMT (see work by Marine Carpuat)
  - Confidence estimation for SMT (see Lucia Specia journal article)
  - Computer-aided Translation (see Koehn tutorial)

- Project: Cross-Lingual Lexical Substitution
  - Cross-lingual lexical substitution is a translation task where you given a full source sentence, a particular (ambiguous) word, and you should pick the correct translation
  - Choose a language pair (probably EN-DE or DE-EN)
  - Download a word aligned corpus from OPUS
  - Pick some ambiguous source words to work on (probably common nouns)
  - Use a classifier to predict the translation given the context

- Project: Predicting case given a sequence of German lemmas
  - Given a German text, run RFTagger (Schmid and Laws) to obtain rich part-of-speech tags
  - Run TreeTagger to obtain lemmas
  - Pick some lemmas which frequently occur in various grammatical cases
  - Build a classifier to predict the correct case, given the sequence of German lemmas as context
  - (see also my EACL 2012 paper)

- Project: Wikification of ambiguous entities
  - Find several disambiguation pages on Wikipedia which disambiguate common nouns, e.g.
    http://en.wikipedia.org/wiki/Cabinet
  - Download texts from the web containing these nouns
  - Annotate the correct disambiguation (i.e., correct Wikipedia page, e.g.

    http://en.wikipedia.org/wiki/Cabinet_(furniture) or (government)
  - Build a classifier to predict the correct disambiguation
    - You can use the unambiguous Wikipedia pages themselves as your only training data, or as additional training data if you annotate enough text

- Project: Moses DE-EN
  - Download and install the open-source Moses SMT system (you may want to use the virtual machine distribution)
  - Download an English/German parallel corpus, e.g., from Opus or statmt.org
  - Build a Moses SMT system for DE to EN
  - Test your system on data from Wikipedia or similar (be sure to check that the English Wikipedia does not contain this content!)
  - Perform an overall error analysis of translation quality
  - Pick some polysemous DE words and show whether Moses can correctly select all of the senses

- Project: Moses EN-DE
  - Download and install the open-source Moses SMT system (you may want to use the virtual machine distribution)
  - Download an English/German parallel corpus, e.g., from Opus or statmt.org
  - Build a Moses SMT system for EN to DE
  - Test your system on English data from the UN multilingual corpus
  - Perform an overall error analysis of translation quality
  - Pick some polysemous EN words and show whether Moses can correctly select all of the senses

- Project: Google Translate X-DE (Pivoting)
  - Select a Language X text for which there is unlikely to be parallel English or German parallel data available (i.e., don't take a classic novel or news!). Suggestion: Wikipedia articles (on topics with no English or German pages)
  - Run this text through Google Translate X-DE
    - Split sentences to be separated by blank lines
    - Carefully save the results and record dates for all translations
  - Explicit pivot
    - Run this text through Google Translate X-EN
    - Post-edit the EN output to fix some obvious major errors
    - Run the original EN output and the post-edited EN through Google EN-DE
  - Perform a careful analysis of Google Translate's performance in translating these texts
    - Is Google Translate "pivoting" when translating from X-DE directly?
    - What are common problems in each translation?
    - Is there useful information which is easier to get from the original X input than from the intermediate EN?
    - Does post-editing the EN help DE translation quality? By how much?

Topics from Fabienne Braune and Matthias Huck

- We are now done with topics (more on Referat/Hausarbeit next)
  - I am also open to your own topic suggestions (should have some similarity to one of these projects)

Referat

- Tentatively (MAY CHANGE!):
  - 25 minutes plus about 15 minutes for discussion
- Start with what the problem is, and why it is interesting to solve it (motivation!)
  - It is often useful to present an example and refer to it several times
- Then go into the details
- If appropriate for your topic, do an analysis
  - Don't forget to address the disadvantages of the approach as well as the advantages
  - Be aware that advantages tend to be what the original authors focused on!
- **List references and recommend further reading**
- **Have a conclusion slide!**

Languages

- I recommend:
- If you do the slides in English, then presentation in English (and Hausarbeit in English)
- If you do the slides in German, then presentation in German (and Hausarbeit in German)
- Additional option (not recommended):
  - English slides, German presentation, English Hausarbeit
  - Very poor idea for non-native speakers of German (you will get tired by the end of the discussion because English and German interfere)

References I

- Please use a standard bibliographic format for your references
  - This includes authors, date, title, venue, like this:
  - (Academic Journal)
  - Alexander Fraser, Helmut Schmid, Richard Farkas, Renjing Wang, Hinrich Schuetze (2013). Knowledge Sources for Constituent Parsing of German, a Morphologically Rich and Less-Configurational Language. *Computational Linguistics*, 39(1), pages 57-85.
  - (Academic Conference)
  - Alexander Fraser, Marion Weller, Aoife Cahill, Fabienne Cap (2012). Modeling Inflection and Word-Formation in SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL),* pages 664-674, Avignon, France, April.

References II

- In the Hausarbeit, use *inline* citations:
  - "As shown by Fraser et al. (2012), the moon does not consist of cheese"
  - "We build upon previous work (Fraser and Marcu 2007; Fraser et al. 2012) by ..."
  - Sometimes it is also appropriate to include a page number (and you *must* include a page number for a quote or graphic)
- Please do not use numbered citations like:
  - "As shown by [1], ..."
  - Numbered citations are useful to save space, otherwise quite annoying

References III

- If you use graphics (or quotes) from a research paper, MAKE SURE THESE ARE CITED ON THE *SAME SLIDE* IN YOUR PRESENTATION!
  - These should be cited in the Hausarbeit in the caption of the graphic
  - Please include a page number so I can find the graphic quickly
- Web pages should also use a standard bibliographic format, particularly including the date when they were downloaded
- I am not allowing Wikipedia as a primary source
  - After looking into it, I no longer believe that Wikipedia is reliable, for most articles there is simply not enough review (mistakes, PR agencies trying to sell particular ideas anonymously, etc.)
- You also cannot use student work (not PhD peer-reviewed) as a primary source

- Any questions?