PhD Thesis Proposal Review

Proposal title: Cross-lingual information retrieval systems
Author: Mgr. Shadi Saleh
Opponent: doc. Ing. Zdeněk Žabokrtský, Ph.D.

A general remark on this version

Most of the text under the current review is identical or very similar to the version submitted in September 2016 (with the exception of a new related work subsection focused on word embeddings, a new subsection on new experiments with document translation, and a new subsection on query expansion). That is why this review is almost identical too. However, my remarks that I find completely resolved in the newer version of the text are now marked using a strikethrough font.

Topic description

The aim of the proposal under review is to develop and evaluate a system that is capable of searching documents relevant for given queries, while the language of the documents is different from the language of the queries.

The proposal consists of six sections. After a short introductory section, a short section which gives a motivation for the task follows. The third section gives a survey of related work. The fourth section mostly describes students's own contribution. The fifth section lists author's publications and the sixth section brings suggestions on future work.

Comments and evaluation

The general motivation behind the proposal is clear. The overall style of the text is quite narrative and thus it seems relatively easy to follow at the first sight. However, after a deeper look one can find some weak (inconsistent or unclear) points in the argumentation that should be probably clarified in the future versions of the text:

- The author claims that his system is a state-of-the-art system (*UPDATE: now page 2*), but actually no other system is evaluated in the text (beating third-party MT systems after plugging them into the presented framework is not enough for an IR state-of-the-art claim).
- The listing of CLIR strategies at the beginning of sec. 3 is not exhaustive: yes, a query can be translated to the documents' language, or documents can be translated to the query's language, but in addition both can be translated (in a wide sense) to some completely different representation too.
- ~~The explanation why translating queries worked worse than translating documents sounds like a common-sense one, but is a little bit speculative in fact (it's rather a post-hoc rationalization, as there can be more possible explanations and no specific evidence is presented).~~
- ~~Related to the previous point: it's surprising that the author explains why translating documents is clearly better (page 3), but then he decides to translate queries.~~
- The semantics of presented evaluation numbers (especially percentages) sometimes remains unclear in the related work section, e.g. in the excerpts from Xu et al. *UPDATE: this still applies, many percentage expressions even in the 4th section are unclear to me.*
- A strict requirement for each PhD thesis is that there should be a very clear cut between author's own contribution and contribution of his colleagues, especially if all student's publication are joint works (having only joint publications is not a big issue in my opinion, just that the author's own contribution must be delimited even more carefully then).

- I'd recommend to restructure (reorder) the presentation of results slightly: I think that the way how meta-parameters were selected (leave-one-out on training data) should be presented prior to the final evaluation using test data. The choice of the HTML stripper should be considered a meta-parameter too (now it is not explained how exactly the choice was made).
- ~~Some claims are presented repetitively (such as the observation that human-better translation does not imply IR-better translation).~~
- The presentation of related work sometimes contains details that are not necessary (e.g., why the "Indri query language" is mentioned as a weighting tool – its function remains totally unclear anyway and it has no implication for the text)
- A lot of attention in the literature survey part is paid to CLIR experiments that were performed almost 20 years ago. Several generations of MT systems appeared in the meantime, and their performance is radically different from that of 1990s, and it is quite likely that those old observations are nowadays not relevant any more.
- Some expressions are simply unclear to me (e.g. ~~"tuned weights" assigned to translations, at the bottom of page 6~~, or "estimating the term translation probabilities from the n-best lists")
- Ad "However, all of the above approaches are based on machine learning approaches, which require large data for training" - It is not clear to me why a contrastive rhetorical figure is used here, since the alternative mentioned below uses large data too, just some other kind of.
- An attention should be paid to results rounding. It is surprising if four of even five significant digits are used if there are just several tens of queries in the test set. Similarly, using four significant digits in Table 4 can be hardly justified.
- The fact that CLIR works often better than monolingual IR using human translated queries deserves some discussion as it is quite unexpected.
- I don't fully agree with the claim that all presented features are independent of the source language. For instance, the distribution of direct and inverse phrase translation probability estimates might be heavily influenced by the degree of variation on the source side (which might be influenced e.g. by the amount of inflection in the source language).
- *UPDATE: The student's CLEF 2016 publication is still marked as accepted for publication only, while it has been published in the meantime.*

Language errors:

- ~~using uppercase letter in inappropriate places "However, Using", "information retrieval Tasks", "source language (Query)"~~
- ~~articles "using an", "by the a Perl module", "in a descending way"~~
- ~~subordinating clauses are incorrectly moved to separated sentences ("Because CLIR systems" page 5, "While for German..." at page 7~~
- ~~an opposite problem: a missing clause/sentence boundary "… by medical experts we ask them", "...we can work… even this might be...", ") This retrieval…"~~
- ~~word order: "the way how reranking implemented is"~~
- ~~"non medical expert" is probably intelligible (if it is to mean a person who is not a medical expert), but formally incorrect~~
- ~~"In the other hand" - "On" should be used~~
- ~~other typos: "It aims to improves",~~ "orgnisers"

Typesetting issues:

- There are (quite regularly) superfluous spaces in front of footnote references (perhaps because of a newline in front of \footnote?). In case of footnote 10 it even leads to a line break.
- Superfluous spaces in front of a comma (page 4, page 7)
- ~~The enumerate LaTeX environment should be perhaps used instead of numbering the paragraphs manually.~~

Conclusion

Shadi Saleh shows that he possesses a good orientation in the field of Information Retrieval, he performed a number of experiments, and some of the innovations led to positive effects on evaluated criteria, which can be considered a promising basis for the future dissertation.

In Prague, 22nd April 2017

Zdeněk Žabokrtský
Institute of Formal and Applied Linguistics
Charles University