Review of thesis proposal by Tom Kocmi: Document Embeddings as a Means of Domain Adaptation for the Machine Translation

Jan Hajič

Prague, Oct. 27, 2017

The thesis proposal first introduces the problem of domain adaptation and the technique of document embeddings. It then summarizes various approaches to domain adaptation to MT. The core of the proposal is Chapter 4, where the author describes his planned approaches to MT domain adaptation by including document embeddings. The thesis then also lists some results already obtained in the area by the author.

In general, it is clear what the direction of the reserrach is and what the research question(s) and subquestions are. It is also apparent that the author has a pretty good familiarity with the field and relevant literature.

What I would expect to be included in more detail, is the core Chapter 4, which describes the plan (it also less than a page long while the thesis has 9 pages of text). It does refer to the previous chapters which describe known approaches and techniques and which the author wants to employ. However, arguments for using (or not using) those earlier approaches or techniques are missing. For example, the text in Chapter 4 does not mention an intention to use synthetic in-domain data (backtranslation) – is there a reason for leaving it out? It looks more of an omission, since the author is a co-author of a paper using this technique as described in Sect. 5.7. In general, the split of plans in Chapter 4 and description of future experiments (Chapter 6) do not help to understand the plans easily. The sequence of plans described in Chapter 6 is however understandable, even though it would be nice to see how possible failures (i.e., results which do not lead to SoA improvement) would lead to change of plans, which inevitably happens.

Another thing to be improved in the thesis that would greatly help to understand it is diagrams and figures. This is natural in using DNN, and the addition of document-level embeddings calls for such diagrams, especially to show the differences in various architectures, for example in Sect. 2.1.

Language-wise, there are some comments in the PDF, but I gave up very soon in marking every language problem with the text, since there is a number of them (from formal ones like missing comma at the end of sentence to wrong prepositions to the use of definite articles. Definite articles are used very often superfluously, even though missing “the”s are also there – please let your texts check by native or near native speakers familiar with the field).

Overall, however, it is clear that the author has deep understanding of the field of NMT in general and adaptation and word/document embeddings use in particular. The author has already published quite a few publications, impressive for this stage of graduate studies. The plans are clear, even if more structure (and “B plans”) would be good to guide the work for the next 2-3 years.