Evaluation report on the half-time report for the Master thesis project proposal “Automating Concept Alignment for Machine Translation” by Arianna Masciolini.

Judgement:

Accept.

General comments:

The work is quite impressive, and it appears you are well on your way to complete a high-quality Masters’ Thesis. Good job!

My main concern is with the writing. Although your English is impeccable, you tend to write rather convoluted sentences. In fact, there are many examples of whole, fairly lengthy, paragraphs consisting of a single sentence. This makes reading your text significantly more difficult than I think you intend. Breaking down lengthy sentences into several simpler sentences would make it much easier for the reader to grasp the concepts you are describing. To give an example, on page 13 you say:

If, as in the case of this project, the aim is to develop a multilingual MT system, these two tasks can be seen as two potentially subsequent steps: in fact, if the first objective is that of obtaining a set of concepts by comparing two or more translations of the same text, once these concepts are in adequate number and of sufficient quality it becomes possible to provide support for additional languages by simply looking for the concrete expressions that, in the new language, correspond to each of the previously gathered concepts.

I have to admit I found that sentence quite a lot to “chew”. If you instead have written something like:

If the aim is to develop a multilingual MT system, these two tasks can be seen as two potentially subsequent steps. In fact, the first objective can be seen as that of obtaining a set of concepts by comparing two or more translations of the same text. With this view, it becomes possible to provide support for additional languages by simply looking for the concrete expressions that correspond to each of the previously gathered concepts.

I think the paragraph had been much easier to follow and (I believe) conveyed the same information.

Similarly, on page 14 you say:

We emphasize the fact that the MT pipeline presented in this thesis and, as a consequence, the GF lexicon we aim to generate are domain-specific since their purpose is not to compete with any existing MT system but rather to provide us with a way to evaluate our CA component, which, being based on syntactic comparison and not, excepts marginally, data-driven, should be able to perform well - in the sense of finding many exact correspondences - on parallel corpora of any size, even on individual sentences.

which, again, is an unnecessarily complex sentence. You do not need to say **everything** in a single sentence!

Finally, you manage to make a 7-line paragraph to be a single sentence on page 26:

Because we want the results of our evaluation of the CE module to be independent both from the previous and the successive stages of the MT pipeline outlined in Section 2.2.3.2, the data we use for this purpose is a subset of the Parallel UD (PUD) corpus, a set of multilingual treebanks in CoNLL-U format

created for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies [2] by means of manual annotation4, which prevents the CE module from failing because of parse errors.

Clearly you could restructure this paragraph, so it does not require an "infinite" reader memory to parse it! For example,

The Parallel UD (PUD) corpus is a set of multilingual treebanks in CoNLL-U format, created for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies [2]. Because we want the results of our evaluation of the CE module to be independent both from the previous and the successive stages of the MT pipeline, we use a subset of the PUD corpus to test our algorithm. This also has the benefit that the CE module will not fail due to parse errors.

Of course, these comments are stylistic and not something you **have** to obey. However, I found the half-time report very time consuming to read since so much of the text required several re-reads to "distill" the relevant information. Again, I would strongly encourage you to work on simplifying your sentence structure and not "cram" in every piece of information in every sentence! You are not charged per sentence!

Minor comments:

Page 7, line 6: "Both formalisms are meant precisely to characterize…" – should be --- "Both formalisms are meant to precisely characterize…"

Page 8, lines 14-16 from bottom: An overly complex sentence. Let me suggest you replace: "Existing dependency parsers, such as UDPipe [33] and the Standford parser [9], are often - but not always [6] - neural pipelines trained on dependency treebanks, significantly more robust than their phrase structure counterparts." with something like: "Existing dependency parsers, such as UDPipe [33] and the Standford parser [9], are often - but not always [6] - neural pipelines trained on dependency treebanks. As a result, such parsers are usually significantly more robust than their phrase structure counterparts."

Page 9, line 5: POS is used before it is defined (at least I could not find it before). As a general rule, I suggest you use the "rule of thumb" that you never use an abbreviation unless it is 1) used throughout the thesis (3-4 at most), or 2) it is defined on the same page. Thus, you may have to spell it out several times, but you make it a lot easier for the reader.

Page 12: line 7 from bottom: "…expressions, one in the source and one in the target language, that are one the translation of the other, which means, from…". Not only complicated, but I could not even comprehend what you intended to say here. Please rewrite!

Page 13 lines 17-23: Way too complex. Please rewrite!

Page 13, line 16 from bottom: "…traditional, even very early MT systems." – should be --- "…traditional, even very early, MT systems."

Page 14, lines 3-6 from bottom. Too complex sentence. Please rephrase!

Page 17, lines 13-16 from bottom: Overly complex sentence. Please break up in smaller sentences!

Page 17, lines 6-12 from bottom: I found your “description” of the algorithm very confusing. May I suggest you provide a few examples to illustrate it? In addition to some pseudo code (which I think you are already planning to add).

Page 17, line 2 from bottom: “This allows to find single-word alignments that would otherwise be ignored.” --- should(?) be --- “This allows it to find single-word alignments that would otherwise be ignored.”

Page 19, lines 6-9 from bottom: Another example of overly complex sentence that I think you can make much easier for the reader to understand if you break it up in smaller sentences.

Page 23, line 13 from below: Question: Are you essentially using weighted voting here? Every alignment approach gets a vote, but the vote is multiplied by some weight to ensure certain alignment approaches have more say in the final result. Or am I missing something? If you are using weighted voting, I think it would be easier to say so.

Page 24, line 16 from bottom: “The roots n_1, tn_2 of two trees…” --- should be --- “The roots n_1, n_2 of two trees…”