PhD Thesis Proposal – Opponent's Review

Thesis Title:	Semi-supervised Machine Learning Methods for Developing Derivational Networks
Thesis author:	Jonáš Vidra
Review author:	Ondřej Dušek
Date:	7 September 2021

Contents summary: The proposed thesis focuses on learning derivational networks, i.e. connecting words of a given language to indicate morphological derivation relations between them. The stated aim of the thesis is to explore methods that require little or no human input or annotations for this task. The practical application focuses mainly on a cross-lingual transfer of human-annotated derivational networks.

The thesis proposal briefly sketches this aim in the introduction, then goes on to describe several complexities and ambiguities in word formation. This is followed by an overview of the author's experiments so far as well as a short guide to the following sections of the proposal. After a related work overview in Sect. 2, the bulk of the text details the author's past experiments. Sect. 3 presents a baseline rule-based cross-lingual transfer based on vocabularies extracted by automatic word alignment. Sect. 4 adds two extensions involving pretrained word embeddings, based on measuring semantic distance or representing semantics of derivational affixes. Sect. 5 represents a different approach, where derivational relations between words are classified using logistic regression with a set of predefined features. All approaches are evaluated and compared to multiple baselines in Sect. 6 and 7. Sect. 8 then briefly details some potential future work extensions and is followed by a short concluding summary.

Overall evaluation: I believe that the thesis topic and aim are novel, interesting and viable. The presented experiments can certainly be a basis of a successful PhD thesis. The methods chosen by the author are mostly plausible and yield positive results to an extent; some shortcomings and problems have been identified by the author and can be improved upon. There are, however, more questions and potential limitations of the work that should preferably be addressed within the thesis (see comments below). The planned future work and extensions are very reasonable, but they feel more like the immediate next steps rather than a comprehensive experiment plan for the remainder of the whole thesis. This should be discussed during the doctoral exam (see also comments below).

High-level comments and questions: Here, I list the most important questions that should be addressed either during the doctoral exam, or in the author's future work on the thesis:

- I believe that the aim of the thesis is a bit too general and should be specified in more detail, perhaps split into multiple sub-aims/research questions/hypotheses.
- The future experiment plan feels a bit like an afterthought – it is not even introduced in Sect. 1.3. Does the future work described in Sect. 8 represent the only experiments planned for the rest of the dissertation work?
- What is the timeframe for the future experiments?
- Are there any alternative paths in case satisfactory results are not reached in future experiments?
- What are the limitations of the approach (or the whole thesis) with respect to language typology and/or the relatedness of languages used for the cross-lingual transfer? All of the experiments only use related Indo-European languages, and some parts of the approach require the use of alphabetic scripts. What happens if we remove these assumptions?
- It would be interesting to see how relatedness of words is perceived by untrained humans, in contrast to linguists. This might be closer to the results of your trained classifiers. Unlike linguists, your models do not have access to historical language development.

- I am not sure if modelling derivations using affixes only, while (mostly) ignoring changes in the stem, is appropriate. Could this be complemented by Levenshtein distance, potentially weighted with respect to word-internal positions of the changes?
- The approaches based on word embeddings mostly use embeddings trained on language modelling tasks. Would a different choice of task lead to a different result? Have you considered e.g. GloVe embeddings, which use a different training objective from all the ones listed in Sect. 4.1? Would it be possible to train embeddings with custom tokenization based on the derivational affixes?
- I believe that using a neural approach in combination with the embeddings should lead to much better results than combining them with rules.
- Do you have any plans on extending the input features for the classification-based approach?
- Would it be possible to use inflectional morphology (PoS tagging or morphological inflection generation) as a source of annotation or as an auxiliary task for training your embeddings or classifiers? I believe that architectures similar to morphological inflection generation (character-based dual encoders or encoder-decoder with monotonic attention) could be usable. The SIGMORPHON shared tasks could be a source of inspiration for this.
- Would it make sense to use neural MT, capable of producing previously unseen words, instead of the alignment dictionaries? This should make translating even rare/theoretical word formations possible.
- A "low-resource" setting is mentioned in Sect. 1.2 but it is not discussed further. I believe that this would be a very interesting type of experiment: starting with a few examples for each relation type and expanding the network to the whole dictionary.

Detailed comments: These serve mainly as feedback for the author. Most of them are focused on ease of reading:

- The extent of related works description is OK, but the descriptions of individual works could could go a bit more to the core/principle of the approaches.
- I thought that in some multilingual word embedding approaches, you can use non-parallel data only? Is this not the main aim of unsupervised machine translation (pg. 3)?
- You should be more explicit regarding the usefulness of dictionaries and translation at the end of pg. 3, it is not obvious why you describe them.
- I am not sure I can agree about the higher specificity of character embeddings compared to word embeddings (pg. 4). Perhaps you need to specify some conditions?
- Is the "translation score" in Sect. 3 the overall score for the derivation?
- Instead of "optimal setting diverges" in Sect. 4, I would say that there is a tradeoff between the two properties.
- I am not sure I understand what you are doing with the similarity scores, why and how exactly you perform the clustering described in Sect. 4.2.
- The difference between the scores in Fig. 1 looks very small. What is the threshold here?
- There is probably no point in showing the right histogram in Fig. 2.
- If you are not the main author of the collaboration papers (Sect. 4.2), you will need to explain your contribution to that work in detail in your thesis.
- You should make clear from the start that the only difference between the "internal" and "external" measures is recall (pg. 8).
- I do not understand why the "upper bounds" (pg. 8) are not included with the oracle score, or why they are not shown in the results tables.
- Are you sure F1 is the correct weighting of precision and recall? You might consider putting more weight on precision. This would obviously need to be explained in any paper or the thesis.

Language/writing: The proposal is written in very good English and is perfectly comprehensible. I just have a few minor notes or suggestions on the writing:

- I was missing an introduction to the task itself; the proposal starts right off with the difficulties. This may be fine in a very focused publication venue, but you would need more explanation in a generic NLP conference or in the thesis itself.
- The individual sections should also have a short introduction that puts them into context (especially when you turn them into thesis chapters). For example, it is not immediately clear from the start of Sect. 3 that this is your own work.
- The proposal is a bit too technical in some places and is hard to digest. The level of detail probably needs to be retained for the thesis or for any future papers, but a bit more high-level description would have been appropriate for the proposal. I would certainly prefer more diagrams over equations, especially over "equations" expressed in prose (pg. 4).
- You should always refer to a specific section number when you refer to other parts of the text (or the experiments contained therein). This is especially visible in Sect. 7, but also in other sections.
- There a few small problems with syntax and articles (e.g., the title of Sect. 1.1 should be "Why word formation is difficult").
- You should not use contractions ("we've", "it's" etc.) in a formal text such as an academic paper or thesis. I might be overly conservative, but this really caught my eye every time.
- The abbreviations in Table 2 and 3 headings should be explained explicitly. I would avoid using "X" to abbreviate "trans", it may lead to confusion.