PhD dissertation – Opponent’s review

Thesis author: Emil Svoboda

Thesis title: Modelling Compounds for Multilingual Data Resources

Opponent’s name: Jiří Hana

Opponent’s affiliation: Charles University, MFF, ÚFAL

The thesis presents

- a detailed in-depth analysis of morphological compounds,
- an enhancement to DeriNet, a Czech database of words related by derivation,
- NLP tools for labeling and analyzing compounds (some for Czech, some multilingual).

The thesis focuses on seven languages, including Czech and English. The dataset and tools are publicly released (under a non-commercial-use license).

This dissertation makes a **substantial contribution to our understanding of language**, both theoretically and computationally. The topic is hard and fuzzy, and the contribution of the author is considerable. I wholeheartedly recommend the thesis for defense.

Below, I list several areas that could be improved, but I want to emphasize that the length of this list is a reflection of the work's significance, not a criticism. This research has the potential to make a real impact, and as such, it's worth taking the time to refine it further in a future publication.

Ease of reading

The text presents a wealth of information, but its presentation could be improved. The reader is left to do a significant amount of work to connect the ideas:

- Most chapters have no summary/conclusion. Larger sections (e.g., 2.1. Definition of compounds) would deserve it too.
- The linguistic part (Sections 2.1 and 2.2) should summarize the view of compounding you adopted/developed. Then, the tools/data sections (in Chapter 3) could modify it for practical reasons. Instead, I had to synthesize the information myself.
- Section 3.1.1 Challenges (of defining and analyzing compounds) should build on Sections 2.1/2.2. Some of it is repetition, and some of it seems to be adjustments made for practical reasons. I had to merge the content of this section with the information in Chapter 2 myself.
- Similarly, Section 4.1.2 Neoclassical compounds repeat verbatim paragraphs from Section 2.1.4. Instead, it should build on that section.
- Presentation of the tools (Sections 3.2, 3.3., 3.4) could be improved:
    - It should be clearer how the tools differ and what properties do they share.
    - Is WFA.ces still useful? Isn't it enough to mention it in a footnote?
    - The section names should be more consistent. For example, evaluation is in "Tool performance", "Performance evaluation and error analysis" and "Experiments and evaluation" (but experiments have a separate section in the other tools).

Minor things

- It would be good to add transliteration for Russian and Greek examples and always used the Roman version when talking about the examples in the text.
- Replace the repeated example *krvotok* (archaic word for bleeding, mostly menstruation; as far as I know, there is no widely shared meaning in current Czech; it does not mean blood flow as suggested) with something more common (maybe *vodopád* 'waterfall' would work?).
- Page 24, example (21): You agree with Scalise and Vogel that the speaker can interpret "hard ball" as "ball which is hard". But isn't it more likely that it is formed in analogy with "football", "basketball", etc.? Similarly, "hardball" and "softball". The interpretation of "football" as "ball that is foot" seems weird. In general, I think that analogy, folk etymology and backformation etc. is often overlooked when analyzing compounds.
- Page 25: You use the term "multi-level" (e.g., Bozdechova's classification, p 25) as synonymous to multi-dimensional. I would prefer the latter expression.
- Several times, you refer to your insights into Czech as "unique". I think you should use "native". Your insights into "French" and "Greek" are also unique.
- Section 3.1.2 A general solution seems out of place.
    - Section 3.1. is labeled Problems and the solutions, 3.1.1 discusses computational approaches to compounding. Considering this, I would assume that the "general solution" would propose a general and flexible approach to compounding. Instead, the section discusses the technical details of deep learning, the evaluation of neural networks, etc.
    - Paragraph Evaluation of neural models (p40).
        - Why is this buried in 3.1., a section mostly discussing linguistic aspects of compounding?
        - There is nothing specific about neural models. This should be called "Evaluation of NLP systems" or something similar.
        - The text mixes general and project-specific topics without any clear indication of which is which. For example, when discussing the concept of training/development/evaluation data, you say that training data is 60% of all the data. That is hardly a general requirement.
        - The selection of terms you define and those assumed to be known seems quite random. For instance, recall and F-measure are defined, but true positives and confusion matrix are not.

Formatting and typography

- The formatting of numbers is inconsistent and sometimes unusual (some instances seem to use nonproportional font). Sometimes, multiple formats are used on the same page without any apparent reason. For example, on page 27, DeriNet versions 2.0 and 2.1 are typeset differently. Just use the same font you use for the surrounding text.
- Example references should follow a consistent format. You use many different formats: ex. (2.3.6), example (38), (ex. 41), (cf. ex. 59), (cf example 51), ...

Questions

- On page 16, last paragraph, you define Czech preposition-like morphemes with two syllables or more as preposition-based roots, and those with fewer syllables as prefixes. What is the rationale behind this distinction, and why is one syllable not sufficient? In Czech prepositional phrases, only non-syllabic prepositions function as clitics. Maybe it would complicate other things?
- You cite Haspelmath (p9), who argues that compounds are not a distinct category. Bauer (p17) makes a similar argument regarding neoclassical compounds. They view these categories as part of a fuzzy, unclustered continuum. You do not explicitly accept or reject their perspective. However, this perspective appears to be at odds with the rest of your approach. Is the reason for not adopting their approach the difficulty of applying it to developing language resources, such as DeriNet and UD corpora? Or is there a fundamental reason why you chose not to adopt their perspective?
- It would be nice if you could highlight any aspects of the definition of compounds that make sense linguistically but need to be modified to make the definition computationally usable.

To summarize, the dissertation presents a substantial amount of work, and I recommend it for defense without reservation.

Prague, September 15, 2024

Jiří Hana