Results of the project PD 125216
Distributional semantics of Hungarian grammatical constructions

The tasks planned for the first year of the project included classification of specific adjunct constructions including identification of their semantics. In the original workplan, I planned to use the HunTag3 chunker to identify NP's in the corpus, but using a dependency parser turned out to be a more accurate alternative: I used Stanford's Graph-based Neural Dependency Parser (the best performing parser in the CONLL 2017 dependency parsing shared task for Hungarian). The parser was trained on the Szeged Dependency Treebank. I performed the semantic classification of the following adverbial adjunct constructions:
- adverbial participles
- negative participles
- adverbs of manner
- specific locative constructions,

identifying the following semantic classes: form of movement, emotion, behaviour, communication, state, amount/estimation, degree, certainty, importance, mental activity, sound emission, preparation, speed, activity, thoroughness, frequency, generality, difficulty, arrangement, fullness, covering. The classification covered 13500 words.

Heads of constructions identified as adverbs of manner and oblique dependents of the predicate of a clause by the dependency parser were clustered using a hierarchical clustering algorithm based on the distributional vectors representing these words in the word embedding model. I evaluated different word embedding models for this task. While models based on a morphologically annotated and lemmatized corpus turned out to be more useful for other tasks (e.g., semantic categorization of nouns in the related FK125217 project), the model built from the original word forms turned out to be better for this task.

I have refined the goal of this and the related FK125217 project in such a way that we aimed at creating a model that can identify the semantics of Hungarian grammatical constructions to the extent that it would be possible to generate sensible questions based on the analysis for the given piece of text. Constituents annotated as "adverbs of manner”, or "oblique dependents" need to be further categorized (as detailed above) so that adequate questions can be generated.

From the point of view of generating questions concerning non-predicative noun phrases (who vs. what), distinctions like person vs. thing or even groups or organizations are relevant. Thus, in the related FK125217 project, we created a semantic categorization of nouns based on these criteria. We also created a verbal argument frame lexicon in which we used thematic roles as the key elements that determine what questions can be generated concerning the predicate itself with the given argument as an anchor. E.g., "What did John do to Jack?" is an appropriate question if John is an agent, and Jack is a patient. In the same situation, "What happened to Jack?" is another appropriate question. We applied the same methodology to phrases containing the instrumental case distinguishing tool usage, sprayed/spread moving patients, agents acting together, having something as a constituent or a feature, using a vehicle, being in a state, wearing something as a garment and various other uses of this case ending.

We categorized
- the 5000 most frequent words in the instrumental case,
- the 2000 most frequent lexical items that have an instrumental dependent.

In the whole process, we used word embedding models to identify and cluster the relevant lexical items. When crating a lexicon of predicate frames, we also distinguished light verb constructions, where it makes no point to ask a question about the nominal element at the core of the construction, e.g., the question "What did you make?" is odd for the sentence "I made a decision." We listed all relevant argument frames of the most frequent 2200 Hungarian verbs, including light verb constructions of 720 verbs.

The first project year was followed by a one-year interruption of both projects due to the birth of my child.

In the following year, the argument frames and roles were algorithmically mapped to occurrences of verbs in the Hungarian Universal Dependencies (HUD) treebank. The model I planned to apply to perform automatic syntactic annotation including thematic role assignment was training a state-of-the art dependency parser on the enriched annotation in HUD. However, the 42000-token (1800-sentence) HUD turned out to be too small to yield the performance I sought. We thus undertook to convert the whole 1.5-million-token Szeged Dependency Treebank (SZDT) to a UD-compatible form for a potential 35-fold increase in training data size. We a) corrected spelling errors (this involved merging/splitting/inserting/deleting tokens), mapped the tokens in the corrected sentences to the original tokenized representations, morphologically analyzed and automatically disambiguated the corrected words, b) reanalyzed inadequately analyzed verb forms, and pronominal elements, c) automatically mapped SZDT dependencies to UD disambiguating one-to-many mappings (e.g. the SZDT ATT relation maps to UD amod, nummod, nmod, ccomp, csubj, acl, advcl, case and subtypes thereof depending on context), and converting head-dependent structures where the two annotation schemes apply different principles. Automatic conversion was performed, this was followed by manual checking of the results. This has not been finished yet but is in progress in the continuing related FK125217 project.

Another goal was investigation of coordinated structures. Coordination is problematic from the point of view of the inherently endocentric dependency paradigm because it is an exocentric construction. Direction of coordination links seems thus to be an arbitrary parameter. In the UD specification, it is always left-to-right. This is most problematic in the case of left elliptic constructions (especially elliptic coordinated NPs), making right-to-left coordination for NPs a preferable solution for Hungarian (also motivated by PPs also being right headed). The lack of case marking in left-elliptic NPs also causes problems for the frame-to-instance mapping algorithm mentioned above if the elliptic left branch is marked as head. The review of dependency parses and parse errors has shown that the handling of coordination is mostly satisfactory, with only sporadic erroneous linking of distributionally dissimilar items (e.g., nouns with different case marking). Similar distribution seems to be a key feature in identifying heads of coordinated constructs. I have found the following (unsurprising) problem areas (nevertheless, most instances of the following constructions are also correctly annotated by the parser):

1. distinction of apposition from coordination (i.e., whether the two ‘conjuncts’ refer to the same entity),

2. distinguishing stacking of locatives of varying granularity like ‘in an office building on Oxford Street in London, England’ from coordination and apposition (due among others to punctuation problems) – consistent annotation of these constructs is far from trivial

3. linking of a mixture of coordinated and subordinated clauses – where to attach what: commonsense knowledge seems to play an important role here,

4. distinguishing subordination and clause coordination in cases where there is no explicit conjunction

5. disambiguation of the scope of coordination and a possessive construction (e.g., Mary and Peter's father).

6. coordination of clauses with verbal and nominal predicates, as well as ellipsis: these are exceptions to the general pattern of coordinated heads having similar distributional properties.

Concerning 6.: the lack of explicit distinction of phrase vs. clause coordination in UD annotation and the lack of copulas in the default present tense 3rd person case makes the interpretation (phrase vs. clause coordination) of even otherwise correct UD annotation difficult in some cases.

Investigation of the semantics of compound structures and identification of bogus compound analyses in morphological analyzer output was the planned topic of my research in this project for the final project year. In the related FK125217project, we managed to achieve very inspiring results with cross-lingual knowledge transfer solving two structured-prediction-type problems: named entity annotation and deep syntactic analysis. In the named entity annotation experiment, we turned a legacy Hungarian named entity resource that distinguished only four entity types into one that covers 29+10 entity types and has 2.8 times as many annotated spans as the original. We trained a new named entity recognizer covering this entity type set. The other zero-shot-transfer-based experiment we performed was evaluating the performance of parser models trained on other languages (English and Czech) generating various meaning representations on Hungarian data. We found that the Prague Tectogrammatical annotation can be mapped to Hungarian with quite acceptable accuracy.

Thus, I took a similar approach to tackle the compound semantics identification problem. I planned to use noun compound semantic relation resources created for other languages and apply the knowledge transfer approach. However, the problem turned out to be much more difficult than I assumed.

There are some resources for English N-N compound semantics. Some of these contain paraphrases of compounds using either prepositions or verbal argument structures. Others contain manual annotation concerning the type of relation between the head and the modifier. The classification of relation types is of varying granularity. Unfortunately, the most sizable resource of this type (Tratz, 2011, about 19000 annotated compounds annotated using 37 relations released as part of the Fanse parser) is no longer available. Another resource by Ó Séaghdha consisting only of 1443 annotated compounds is available. Ó Séaghdha's annotation consists of a hierarchical system of semantic relation types with six coarse relations at the top: BE, HAVE, IN, ACTOR, INST and ABOUT. I also found and downloaded another dataset consisting of Afrikaans and Dutch compounds (AUCOPRO), a few thousand items each that were annotated using Ó Séaghdha's annotation scheme by two or three annotators. Another resource for German (de-nncom-sem, about 8000 compounds) using a different richer annotation scheme is also available.

The Afrikaans and Dutch data explicitly exposes a problem with this style of annotation: low inter-annotator agreement concerning the semantic relations between the compound head and modifier. In the Afrikaans NN compound dataset, only 47% of the compounds received the same annotation from two of the three annotators, and only 28% got an identical annotation from all three. When

doing my experiments, I only used items from the Dutch and Afrikaans data where at least two annotators agreed. However, this reduced the dataset to a fraction of its already very limited size.

I applied a model for compound classification similar in architecture to the one applied to named entity recognition based on the multilingual XLM-RoBERTa language model. The monolingual performance of this model (tested on test sets split from the mentioned data sets) is inferior to the best static-embedding-based models (we obtained around 50% accuracy for all models). This could be due to the fact that the contextual model was trained on sentences rather than separate words. Unfortunately, state-of-the-art results themselves are also quite low (about 60% accuracy for just the top 6 categories of the Ó Séaghdha dataset), because the datasets are too small, and often more than one relation is applicable to the same compound (even at the coarse-grained top level, cf. the IAA on the Afrikaans data), which results in inconsistencies in annotation.

Zero-shot application of the models to Hungarian resulted in further degradation of performance, which was unfortunately more substantial than in the case of the transfer of the named entity recognition model.

For the identification of bogus compound analyses in morphological analyzer output, I created various word-embedding-based models. The best models could identify whether a productive compound analysis by the morphological analyzer is semantically valid or not with 90% accuracy, and 50% recall on bogus analyses (these are much rarer than valid analyses (8% of all)).

Lexical resources created in the project will be made available at nlpg.itk.ppke.hu.