Bachelor thesis exposé

Recognizing and resolving coordination ellipses in German medical discharge summaries

Alexandra Tichauer

9. Januar 2019

Gutachter: Prof. Ulf Leser
Prof. Anke Lüdeling

Betreuerin: Dr. Madeleine Kittner

Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Informatik

1 Motivation

As part of an ongoing project, a research group at the HU Berlin chair of Wissensmanagement in der Bioinformatik deals with the recognition and normalization of named entities, namely diseases, treatments and medications, in German medical discharge summaries. As those summaries are written to efficiently communicate rather large quantities of information between healthcare professionals, they contain many abbrevations, enumerations and elliptical coordinations. The latter pose an obstacle to successfully recognizing named entities, because typical NER systems are not aware of the underlying semantic and syntactic structure of coordinations and may thus process them incorrectly (cf. Blake and Rindflesch [1] p. 1, Buyko et al. [2] p. 1) . To tackle this loss of information, in this work we aim to detect and normalize coordination ellipses in our corpus as a preprocessing procedure for NER.

2 Research Objective

In natural languages, two or more phrases of the same type can be connected through coordination. If they contain lexically identical parts, those can be omitted in all but one of the phrases, as the hearer or reader will be able to infer them from the rest of the utterance. For instance, instead of writing *hepatische Filiae und pulmonale Filiae* (example 1), the author can shorten the phrase to *hepatische und pulmonale Filiae* without changing its meaning. This omission is called **coordination ellipsis** (CE), and is further illustrated by examples 2 and 3 from our corpus.

(1) hepatische und pulmonale Filiae
hepatic and pulmonary metastasis

(2) lymphatische, hepatische, ossaere, zerebrale Metastasen
lymphatic, hepatic, osseous, cerebral metastases

(3) Ausbreitung: hepatisch, lymphatisch
spreading: hepatic, lymphatic

We will use the following terminology to refer to the different parts of elliptical entity mentions (consistent with Buyko et al. [2]):

The **conjunction** is linking the two or more **conjuncts** in the phrase, e.g. *and* in example 1. It need not be a word, but can also be covert, usually represented by a comma (examples 2, 3). The **antecedent** is the part (or parts) of the elliptical mention shared by both conjuncts, like *Filiae* in example 1. A conjunct doesn't include the antecedent. The phrases constructed by combining the antecedent with each of the conjuncts separately will be called **resolved conjuncts**. (In example 1, the resolved conjuncts would be *hepatische Filiae* and *pulmonale Filiae*.)

In contrast to human readers, a NER system will not automatically be able to reconstruct both (or all) resolved conjuncts from an elliptical mention. It may either only recognize the closest conjunct as belonging to the antecedent (it might, for example, only detect *pulmonale Filiae* in example 1). Or it may claim the phrase as a whole to be a single named entity although in fact it contains multiple entities. It is evident that a lot of information will be lost if named entity recognition is carried out without paying attention to coordination ellipses resolution. Thus, our objective is to improve overall NER performance by accomplishing the two following subtasks:

a. detect CE in the text *and*

b. resolve them, i.e. insert the omitted parts and return a non-elliptical version of each document.

3 Related Work

To our knowledge, there are no publications about the detection and resolution of coordination ellipses in a German language corpus. For English however, and especially for the biomedical domain, research has been carried out employing diverse approaches. There is a variety of tools and publications generally dealing with named entity recognition in medical texts (discharge summaries, clinical notes, etc.), for instance cTAKES (Savova et al. [3]), HITEx (Zeng et al. [4]) and MetaMap (Aronson and Lang [5]), although none of them explicitly addresses CE-resolution. It has to be noted, therefore, that all of the subsequently mentioned projects used corpora of biomedical scientific papers or abstracts, such as GENIA [6], PennBioIE [7] and CRAFT [8], as opposed to writings from medical practice.

Buyko et al. [2] search for possible coordination ellipses after the actual NER. Text chunks marked as named entities by their system and containing one of a set of frequent coordinations (e.g. *and, or*) will be considered a candidate entity mention. They then use conditional random fields for sequencing, taking into account features like the token stems, the part of speech tags and the token position, thus distuingishing the conjunction and conjuncts. Finally, ellipsis resolution is achieved by copying the antecedent (tokens neither marked as conjunction nor as conjunct) to every identified conjunct in the elliptical mention. Wei et al. [9] pursue a very similar approach. They however also target range mentions (e.g. *SMAD 2 to 4*) and other types of elliptical mentions, and include certain additional morphological (e.g. suffixes typical for chemical substances) and semantic (e.g. chemical elements) characteristics in the feature set of their CRF-classifier.

Philip Ogren [10] deals with coordination resolution independently from inserting elided material. (However, once the right coordination structure is found, said insertion can be done with very high accuracy, as he himself remarks ([10], p. 9)). He compares two main approaches for automatically assigning structures to coordinated sentence constituents.

The first is an algorithm built upon a syntactic (dependency tree) parse of his corpora and consisting of manually developed rules. The second relies on machine learning algorithms to classify tokens to the left and right of a conjunction with a binary distinction, i.e. identify them as being a conjunct boundary or not. While Ogren extensively explores the influence of using different POS-tagsets and machine learning paradigms on overall performance, some of the rather basic ideas pertaining to his work may be more influential for our project. One of them is his „Coordination Structure Production Algorithm“, working on a syntactic parse of a corpus. Further, he incorporates a language model and even orthographic similarity to assess the likeliness of two phrases being conjuncts.

In contrast to the previously described approaches, Chae et al. [11] and Blake and Rindflesch [1] do not rely on machine learning algorithms for the core of their ellipses resolution systems. Instead they identify candidate resolved conjuncts through heuristic rules and try to validate their correctness, using a dictionary and, again, some handwritten rules. Both of them focus on coordinated noun phrases, as their work is meant as an improvement for NER systems. Blake and Rindflesch start from a dependency parse of their corpus. They find candidate elliptical entity mentions via the dependency labels of their tokens (typically adverbial modifier, noun phrase and conjunction), and expand those mentions following a straightforward scheme, similar to the one used by Buyko et al. after sequencing. An interesting detail of their semantic validation module is that they build a dictionary from the very corpora they are working on, calculating overall cooccurence rates for conjuncts and antecedents. This is, indeed, a promising strategy for domains lacking an extensive dictionary. In our case, although dictionaries like the ICD10 are available, actual entity mentions often differ from the dictionary entries in using variable wordings, so that a corpus-intern lexical resource may be an advantage.

4 Data

The corpus to be studied in this thesis consists of 200 German medical discharge summaries, half of them concerning patients with liver cancer, the other half concerning patients with melanoma. All documents have been tokenized and POS-tagged automatically by the JPOS-tagger from the JCORE package (Hellrich et al. [12]), producing app. 230 000 tokens. They are currently annotated by medical experts according to the categories Diagnosis, Treatment, Medication, Localisation, LevelOfTruth, in order to obtain a gold standard corpus. Annotation seeks to relate the found named entities to international standard categorizations, i.e. ICD10 for diagnoses, OPS for treatments and ATC for medications. However, only certain parts of the documents (marked by specific headlines) are considered.

A first survey of a selection of sample documents showed that the occuring coordination ellipses vary greatly with respect to their syntactic structure, length and complexity. At the same time, the corpus is rather small, and therefore the overall number of ellipses will

be limited to an estimated 500 – 600, of which only approximately one third will later be of interest for named entity recognition. The structure and contents of the documents are quite particular - they are written in a telegram-like, short style and contain many abbrevations and enumerations. The vocabulary is, of course, domain-specific and highly repetitive between documents.

5 Approach

Due to the peculiarities of the corpus and the apparent lack of other research on the topic with German texts, we will transfer and adapt approaches from other languages and domains to fit our context. We will limit the CE-resolution to coordinated nominal phrases, because in our case named entities are most likely to be attributed nominal phrases (see examples 1 -3). For other types of coordinated phrases, there are hardly enough occurences involving named entities to allow for the development of robust recognition and resolution rules. One significant part of this work is to manually annotate all nominal elliptical coordinations in the whole corpus (200 documents). This will be done using the on-line tool Brat [13]. The annotated corpus will consequently be used as a gold standard for measuring the perfomance of our CE-resolution algorithm.

The small size of the corpus and relative scarceness and diversity of CE make machine learning approaches unlikely to produce satisfying results. We will therefore base our system on heuristic rules. A set of simple POS-tag sequences typical for coordination ellipses in the corpus (e.g. ADJA CONJ ADJA NN, corresponding to phrases like *hepatic and pulmonary metastasis*) can serve as a baseline for development and performance assessment. It would be desirable to start from a dependency parse of the corpus, using an already available tool. Kara et al. [14] propose a dependency parser specifically designed for German clinical texts, which should exactly fit our domain, and which they show to outperform the Stanford NLP dependency parser for German. The parser is available at the DFKI website [1]. We could then pursue the method described by Blake and Rindflesch [5], i.e. generate possible resolved conjuncts and validate them afterwards. To assess the correctness of a resolved conjunct, in a first step we would try to find it in the ICD10, OPS and ATC dictionaries. Further, we would resort to statistics established from our own corpus. These would contain all non-elliptical noun phrases appearing in the documents, cooccurence rates for specific nominal heads and modifiers and usage counts (general, number of articles) for heads and modifiers. It may be, however, that the proposed dependency parser fails to correctly process the shortened and incomplete sentences of our corpus, and it is improbable that any other less specialized parser could outperform it. In this case, we will have to rely entirely on POS-patterns for finding coordination ellipses (especially for determining their limits) and generating resolved conjuncts. We would then attempt to maximize recall by creating rather transmissive detection rules

[1] `http://macss.dfki.de/dependency_parser.html`

and in turn lay a special focus on validation of resolved conjunct candidates, to eliminate false positives.

The corpus will be divided into a development and a test set. The latter will be annotated only after completion of the algorithm design, to avoid any influence on the establishment of rules (as the annotator and developer will be the same person). There are two key indicators to be considered in evaluation, corresponding to the two steps of the algorithm mentioned above: a. the precision and recall of finding coordination ellipses and b. the accuracy of resolving them. Different categories of errors may occur during ellipsis recognition and resolution: non-elliptical structures could be erroneously treated as ellipses (false positives), elliptical structures may be either overlooked entirely or rejected because their resolved conjuncts can't be validated (false negatives) and found ellipses could be resolved incorrectly. We will analyze both quantity and quality of these errors and explore their main causes, to permit future improvement of the algorithm and provide insights for the development of similar projects.

References

[1] Catherine Blake and Tom Rindflesch. Leveraging syntax to better capture the semantics of elliptical coordinated compound noun phrases. *Journal of Biomedical Informatics*, 72:120 – 131, 2017.

[2] Ekaterina Buyko, Katrin Tomanek, and Udo Hahn. Resolution of coordination ellipses in complex biological named entity mentions using conditional random fields. *ISMB BioLink SIG*, pages 163–171, 2007.

[3] Guergana K. Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[4] Qing T. Zeng, Sergey Goryachev, Scott Weiss, Margarita Sordo, Shawn N. Murphy, and Ross Lazarus. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making*, 6(1):30, Jul 2006.

[5] Alan R. Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[6] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl1):i180–i182, 2003.

[7] Mark Liberman, Mark Mandel, and Peter White. Pennbioie oncology 1.0. LDC2008T21, Philadelphia: Linguistic Data Consortium, 2008.

[8] Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A. Baumgartner, K. Bretonnel Cohen, Karin Verspoor, Judith A. Blake, and Lawrence E. Hunter. Concept annotation in the craft corpus. *BMC Bioinformatics*, 13(1):161, Jul 2012.

[9] C. Wei, R. Leaman, and Z. Lu. Simconcept: A hybrid approach for simplifying composite named entities in biomedical text. *IEEE Journal of Biomedical and Health Informatics*, 19(4):1385–1391, 2015.

[10] Philip Victor Ogren. *Coordination Resolution in Biomedical Texts*. PhD thesis, University of Colorado, Boulder, 2011.

[11] Jeongmin Chae, Younghee Jung, Taemin Lee, Soonyoung Jung, Chan Huh, Gilhan Kim, Hyeoncheol Kim, and Heungbum Oh. Identifying non-elliptical entity mentions in a coordinated np with ellipses. *Journal of Biomedical Informatics*, 47:139 – 152, 2014.

[12] Johannes Hellrich, Franz Matthies, Erik Faessler, and Udo Hahn. Sharing models and tools for processing german clinical texts. *Studies in health technology and informatics*, 210:734–8, 2015.

[13] Pontus Stenetorp, Sampo Pyysalo, and Goran Topić. brat rapid annotation tool, 2018.

[14] Elif Kara, Tatjana Zeen, Aleksandra Gabryszak, Klemens Budde, Danilo Schmidt, and Roland Roller. A domain-adapted dependency parser for german clinical text. In *Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)*. KONVENS, 9 2018.