Effect of Out Of Vocabulary terms on inferring eligibility criteria for a retrospective study in Hebrew EHR

Raphael Cohen*
Computer Science Dept.
Ben-Gurion University in the Negev
cohenrap@bgu.ac.il

Michael Elhadad
Computer Science Dept.
Ben-Gurion University in the Negev
elhadad@cs.bgu.ac.il

1 Background

The Electronic Health Record (EHR) contains information useful for clinical, epidemiological and genetic studies. This information of patient symptoms, history, medication and treatment is not completely captured in the structured part of the EHR but is often found in the form of free-text narrative.

A major obstacle for clinical studies is finding patients that fit the eligibility criteria of the study. Using EHR in order to automatically identify relevant cohorts can help speed up both clinical trials and retrospective studies (Restificar, Korkontzelos et al. 2013).

While the clinical criteria for inclusion and exclusion from the study are explicitly stated in most studies, automating the process using the EHR database of the hospital is often impossible as the structured part of the database (age, gender, ICD9/10 medical codes, etc.') rarely covers all of the criteria.

Many resources such as UMLS (Bodenreider 2004), cTakes (Savova, Masanz et al. 2010), MetaMap (Aronson and Lang 2010) and recently richly annotated corpora and treebanks (Albright, Lanfranchi et al. 2013) are available for processing and representing medical texts in English. Resource poor languages, however, suffer from lack in NLP tools and medical resources. Dictionaries exhaustively mapping medical terms to the UMLS medical meta-thesaurus are only available in a limited number of languages besides English. NLP annotation tools, when they exist for resource poor languages, suffer from heavy loss of accuracy when used outside the domain on which they were trained, as is well documented for English (Tsuruoka, Tateishi et al. 2005; Tateisi, Tsuruoka et al. 2006).

In this work we focus on the problem of classifying patient eligibility for inclusion in retrospective study of the epidemiology of epilepsy in Southern Israel. Israel has a centralized structure of medical services which include advanced EHR systems. However, the free text sections of these EHR are written in Hebrew, a resource poor language in both NLP tools and hand-crafted medical vocabularies.

Epilepsy is a common chronic neurologic disorder characterized by seizures. These seizures are transient signs and/or symptoms of abnormal, excessive, or hyper synchronous neuronal activity in the brain. Epilepsy is one of the most common of the serious neurological disorders (Hirtz, Thurman et al. 2007).

2 Corpus

We collected a corpus of patient notes from the Pediatric Epilepsy Unit, an outpatient clinic for neurology problems, not limited to epilepsy, in Soroka Hospital. This clinic is the only available pediatric neurology clinic in southern Israel and at the time of the study was staffed by a single expert serving approximately 225,000 children. The clinical corpus spans 894 visits to the Children Epilepsy Unit which occurred in 2009 by 516 unique patients. The corpus contains 226K tokens / 12K unique tokens.

∗Supported by the Lynn and William Frankel Center for Computer Sciences, Ben Gurion University

Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013), pages 116–119,
Sofia, Bulgaria, August 4-9 2013.

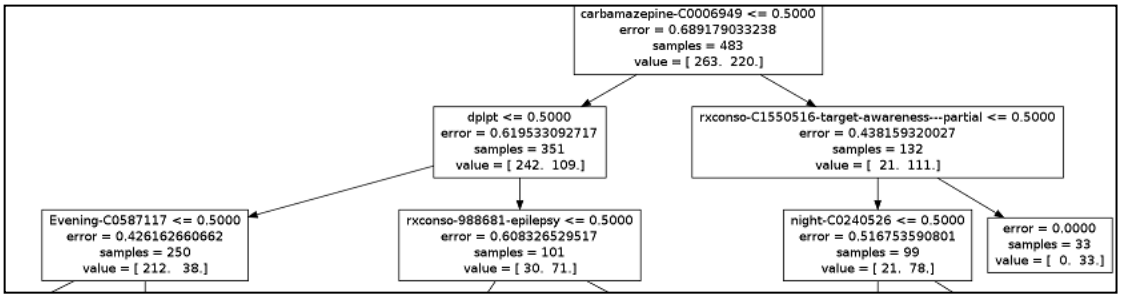

Figure 1 – Decision Tree for inclusion/exclusion. Sodium Valproate (dplpt) is a key term which is often segmented incorrectly.

The patients were marked by the attending physician as positive or negative for epilepsy. In the study year, 2009, 208 patients were marked as positive examples and 292 as negative. The inclusion criteria were defined as history of more than one convulsive episode excluding febrile seizures. In practice, the decision for inclusion was more complex as some types of febrile seizure syndromes are considered a type of epilepsy while some patients with convulsion were excluded from the study for various reasons.

3 Method

We developed a system to classify EHR notes in Hebrew into "epilepsy" / "non-epilepsy" classes, so that they can later be reviewed by a physician as eligible candidates into a cohort. The system analyzes the Hebrew text into relevant tokens by applying morphological analysis and word segmentations, Hebrew words are then semi-automatically aligned to the UMLS vocabulary. The most important tagged Hebrew words are then used as features fed to a statistical document classification system. We evaluate the performance of the system on our corpus, and measure the impact of Hebrew text analysis in improving the performance for patient classification.

4 Out-Of-Vocabulary Terms

The complex rules of Hebrew word formation make word segmentation the first challenge of any NLP pipeline in Hebrew. Agglutination of function words leads to high ambiguity in Hebrew (Adler and Elhadad 2006). To perform word segmentation, Adler and Elhadad (Adler and Elhadad 2006) combine segmentation and morpheme tagging using an HMM model over a lattice of possible segmentations. This learning method uses a lexicon to find all possible segmentations for all tokens and chooses the most likely one according to POS sequences. Unknown words, a class to which most borrowed medical terms belong, are segmented in all possible ways (there are over 150 possible prefixes and suffixes in Hebrew) and the most likely form is chosen using the context within the same sentence. Beyond word segmentation, the rich morphological nature of Hebrew makes POS tagging more complex with 2.4 possible tags per token on average, compared to 1.4 for English.

Out of 12K token types in the corpus 3.9K (30%) were not found in the lexicon used by the Morphological Disambiguator compared to only 7.5% in the Newswire domain. A sample of 2K unknown token was manually annotated as: transliteration, misspelling and Hebrew words missing in the lexicon. Transliterated terms made up most of the unknown tokens (71.5%) while the rest were misspelled words (16%) and words missing from the lexicon (13.5%).

Error analysis of the Morphological Disambiguator in the medical domain corpora shows that in the medical domain, Adler *et al*'s unknown model still performs well: 80% of the unknown tokens were still analyzed correctly. However, 88.5% of the segmentation errors were found in unknown tokens. Moreover, the transliterated words are mostly medical terms important for understanding the text.

5 Acquiring a Transliterations Lexicon

As transliterations account for a substantial amount of the errors and are usually medical terms, therefore of interest, we aim to automatically create a dictionary mapping transliterations in our target corpus to a terminology or vocabulary in the source language. In our case, the source language is medical English which is a mix of English and medical terms from Latin as represented by the UMLS vocabulary.

The dictionary construction algorithm is based on two methods: noisy transliteration of the medical English terms from the UMLS to Hebrew forms (producing all the forms an English terms may be written in Hebrew, see (Kirschenbaum and Wintner 2009)) and matching the generated

transliterations to the unknown Hebrew forms found in our target corpus. After creating a list of candidate pairs (Hebrew form found in the corpus and transliterated UMLS concept), we filter the results to create an accurate dictionary using various heuristic measures.

The produced lexicon contained 2,507 transliterated lemmas with precision of 75%. The acquired lexicon reduced segmentation errors by 50%.

6 Experiments

6.1 Experimental Settings

An SVM classifier was trained using the 200 most common nouns as features. The noun lemmas were extracted with the morphological disambiguator in two settings: naïve setting using the newswire lexicon and an adapted setting using the acquired lexicon.

We divided the corpus into training and testing sets of equal size, we report on the average results or 10 different divisions of the data.

6.2 Results

The classifier using the baseline lexicon achieved an average F-Score of 83.6%. With the extended in-domain transliterations lexicon the classifier achieves F-Score of 87%, an error reduction of 20%.

We repeated the experiment with decision trees for visualization for error analysis. With decision trees we see an improvement from 76.8% to 82.6% F-score. In Figure 1, we see in the resulting decision tree the most commonly prescribed medication for epilepsy patients, Sodium Valproate "*depalept*" ("דפלפט"). This word appears in three forms: "*depalept*", "*b+deplapet*" and "*h+depalept*". The acquired lexicon allows better segmentation of this word thus removing noise for documents containing the agglutinated forms.

7 Conclusions

We presented the task of classifying patients' Hebrew free text EHR for inclusion/exclusion from a prospective study. Transliterated tokens are an important feature in medical texts. In languages with compound tokens this is likely to lead to segmentation errors.

Using a lexicon adapted for the domain impacts the number of segmentation errors, this error reduction translates into further improvements when using these data for down the line applications such as classification.

Creating domain adaptation methods for resource-poor languages can positively impact the use of clinical records in these languages.

Acknowledgments

Adler, M. and M. Elhadad (2006). An unsupervised morpheme-based hmm for hebrew morphological disambiguation. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Association for Computational Linguistics.

Albright, D., A. Lanfranchi, et al. (2013). "Towards comprehensive syntactic and semantic annotations of the clinical narrative." Journal of the American Medical Informatics Association.

Aronson, A. R. and F. M. Lang (2010). "An overview of MetaMap: historical perspective and recent advances." Journal of the American Medical Informatics Association 17(3): 229-236.

Bodenreider, O. (2004). "The unified medical language system (UMLS): integrating biomedical terminology." Nucleic Acids Research 32(Database Issue): D267.

Hirtz, D., D. Thurman, et al. (2007). "How common are the "common" neurologic disorders?" Neurology 68(5): 326-337.

Kirschenbaum, A. and S. Wintner (2009). Lightly supervised transliteration for machine translation. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics.

Restificar, A., I. Korkontzelos, et al. (2013). "A method for discovering and inferring appropriate eligibility criteria in clinical trial protocols without labeled data." BMC Medical Informatics and Decision Making 13(Suppl 1): S6.

Savova, G. K., J. J. Masanz, et al. (2010). "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.

Tateisi, Y., Y. Tsuruoka, et al. (2006). Subdomain adaptation of a POS tagger with a small corpus. Proceedings of the Workshop on

Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis, Association for Computational Linguistics.

Tsuruoka, Y., Y. Tateishi, et al. (2005). "Developing a robust part-of-speech tagger for biomedical text." Advances in informatics: 382-392.