The JULIE LAB MANTRA System for the CLEF-ER 2013 Challenge

Johannes Hellrich and Udo Hahn

Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena
Jena, Germany

johannes.hellrich@uni-jena.de
udo.hahn@uni-jena.de

Abstract. We here describe the set-up for the system from the Jena University Language & Information Engineering (JULIE) Lab which participated in the CLEF-ER 2013 Challenge. The task of this challenge was to identify hitherto unknown translation equivalents for biomedical terms from several parallel text corpora. The languages being covered are English, German, French, Spanish and Dutch. Our translation system enhanced, in a realistic scenario, the French, German, Spanish and Dutch parts of a UMLS-derived terminology with 4k to 15k new entries each. Based on expert assessment of the new German translations about 76% of these were judged as plausible term enhancements.

1 Introduction

The task underlying the CLEF-ER 2013 Challenge is to identify translation equivalents for biomedical terms from several parallel text corpora. Terms from this sublanguage are either single-word (e.g. *'appendicitis'*) or multi-word terms (e.g. *'blood cell'*). The task not only encompasses the recognition of literal mentions of these terms but also includes a grounding task, namely to determine unique identifiers for the recognized terms in an authoritative biomedical terminology, the Unified Medical Language System (UMLS).[1]

The UMLS is an umbrella system for a huge collection of various specialized domain terminologies for human anatomy, diseases, drugs, clinical care, etc. that was originally developed for the English language but has subsequently been complemented by translations into a large variety of other languages. Whereas the English UMLS is a rather complete collection of biomedical terms, the non-English versions lack this property to different degrees though. Hence, the proper enhancement of the lexical and conceptual coverage of the non-English UMLSes constitutes a rewarding goal which was picked up by the CLEF-ER 2013 Challenge organizers[2] for the following languages: English, German, French, Spanish and Dutch.

From the perspective of biomedical natural language processing the task combines aspects of both named entity recognition (NER) and machine translation

[1] http://www.nlm.nih.gov/research/umls/

[2] https://sites.google.com/site/mantraeu/clef-er-challenge

(MT). Our system leans more towards the MT side, since we try to solve the challenge by enriching a UMLS-derived terminology with new translations extracted from the parallel texts. This enriched terminology is thereafter used to annotate the corpora with a gazetteer. This eases both linking new entities with the identifiers used in the terminology and a human review of the system's output, since we produce some thousands of new terms to review instead of hundreds of thousands of annotations.

Our terminology translation system works by combining phrase-based statistical machine translation (SMT) with named entity recognition. For this task, we exploit the MEDLINE and EMEA biomedical parallel corpora (see details below) for all relevant language pairs. Our translation system enhanced, in a realistic scenario, the French, German, Spanish and Dutch parts of the UMLS-derived terminology with 4k to 15k new entries each. Based on expert assessment of the new German translations about 76% of these were judged as plausible term enhancements.

2 JULIE Lab's MANTRA System

This section describes the translation part of JULIE Lab's MANTRA system, where we distinguish preparatory steps (Section 2.1) from the candidate generation (Section 2.2) and candidate filtering steps (Section 2.3) and, finally, turn to the results of applying our system to the challenge data (Section 2.4).

A major design requirement during system development was to reuse existing and widely used software, often developed in other contexts, and keep the system as domain- and language-independent as possible. Accordingly, we equipped JULIE Lab's MANTRA with the LINGPIPE[3] gazetteer, GIZA++ and MOSES[4] [1, 2] for phrase-based SMT, JCORE for biomedical NER [3] and, finally, WEKA[5] [4] for learning a maximum-entropy model to combine NER and SMT information.

2.1 Preparatory Steps

To generate training data for the translation part of our system we merged the MEDLINE®[6] and EMEA [5] parallel corpora provided for the CLEF-ER 2013 Challenge, resulting in one file per language pair. These files were then annotated for those biomedical entities already contained in the UMLS-derived CLEF-ER terminology,[7] using a LINGPIPE-based gazetteer. 10% of each corpus were taken apart and used to train the JCORE NER engine. The remaining 90% of the corpora were used to train a phrase-based SMT model with GIZA++ and MOSES.

[3] http://alias-i.com/lingpipe/

[4] http://www.statmt.org/moses/

[5] http://www.cs.waikato.ac.nz/~ml/weka/

[6] http://mbr.nlm.nih.gov/Download/

[7] For technical and data consistency purposes, the original UMLS terminology was slightly curated and reformatted. This specific UMLS version is available at https://sites.google.com/site/mantraeu/terminology

2.2 Candidate production

The model created by GIZA++ contains phrase pairs, such as *'an indication of tubal cancer'* $\rightarrow$ *'als Hinweis auf ein Tubenkarzinom'*, and several translation probabilities for each pair, namely inverse phrase translation probability $\phi(f|e)$, inverse lexical weighting $lex(f|e)$, direct phrase translation probability $\phi(e|f)$ and direct lexical weighting $lex(e|f)$. The probabilities describe conditional probabilities for a phrase in the target language *e* (either German, French, Spanish or Dutch) being a translation of a phrase in the source language *f* (English). Translation candidates for enriching the UMLS were produced by filtering phrasal pairs such that only those were retained which translated a known English synonym to a biomedical concept in one of the target languages.

2.3 Candidate Filtering

We filtered the candidates produced in the previous step by using WEKA to train a maximum-entropy model as a selector that keeps or discards tentative translations. We used as features the phrase probabilities from the SMT model, the NER system's judgment for each candidate being a recognized named entity (thus removing biomedical non-terms) and, for our final submission only, the ratio between the respective character lengths of the translated synonym and its translation equivalent. This ratio was logarithmized to keep features on a similar scale. The NER system's judgment was normalized over all sentences containing the phrase in question by summing the probabilities for it being a named entity and dividing by the total number of sentences, taking '0', if no match was found:

$$p_{NE} = \frac{\sum_{i=0}^{sentences} EntityProbability_i}{sentences}$$

The annotations for the already known translations, generated in step 2.1 were used as training material.

2.4 Results & Evaluation

Those translation candidates accepted by the filter and not yet contained in the UMLS were added to the dictionary used by our gazetteer system. We generated two new dictionaries, one with and one without the length ratio as a feature, and used these to annotate our submissions.

To evaluate our translation subsystem prior to submission, we used both expert judgment and an automatic benchmark, dealing with *new* and already *known* terminology, respectively. In the first one, translations not yet contained in the UMLS were judged by a biomedical expert.[8] In the second setting, we measured the system's performance in recreating portions of the already available versions of the UMLS, i.e. we matched suggested translations with entries already contained in the terminology. This evaluation was done concept-wise, i.e. a word which is a synonym for multiple concepts or a translation thereof was examined multiple times. To measure the effects of different system configurations we used precision, recall and F_1-score calculated as follows:

[8] A bioinformatics graduate student utilizing online medical dictionaries.

$$precision = \frac{correct\ translations}{proposed\ translation}$$
$$recall = \frac{correct\ translations}{traceable\ translations}$$
$$F_1\text{-}score = 2\frac{precision*recall}{precision+recall}$$

A translation was counted as *correct*, if the translation was contained in the set of synonyms in the respective language for the UMLS concept. A translation was considered as *traceable*, if the corresponding concept was annotated in two parallel sentences by the gazetteer system. We evaluated the translation of each synonym of a UMLS entry independently. These criteria aim to measure how much of the UMLS which could have been reconstructed from the parallel corpus was actually reconstructed by our system.
Expert judgment was collected based on a sample of 100 English-German term translations created with all features except the length ratio. 76% of these new translations were judged as being correct and reasonable from a domain knowledge perspective. Automatic analysis was performed for all languages, results are listed in the following table, with a baseline system performing no filtering of term candidates listed for comparison. Both configurations of the JULIE Lab's

Language	Measurement	Baseline	JULIE LAB'S MANTRA system	
			without length ratio	with length ratio
French	F_1-score	0.13	**0.61**	0.57
	Precision	0.07	0.58	**0.61**
	Recall	**0.98**	0.64	0.53
German	F_1-score	0.14	**0.66**	0.65
	Precision	0.07	0.61	**0.69**
	Recall	**0.98**	0.73	0.60
Spanish	F_1-score	0.19	0.72	**0.73**
	Precision	0.11	0.62	**0.63**
	Recall	**0.97**	0.85	0.86
Dutch	F_1-score	0.15	0.79	**0.81**
	Precision	0.08	0.70	**0.76**
	Recall	**0.98**	0.89	0.87

Table 1. Evaluation by partial UMLS recreation. We list results comparing our submissions with a baseline system without candidate filtering.

MANTRA system are clearly superior to the baseline in all regards except recall. The system with length ratio provides similar F_1-scores to the one without, yet we suppose it to be more adequate for the CLEF-ER challenge, due to its higher precision.

3 Related work

Prior efforts to use parallel corpora for terminology translation were performed by Déjean *et al.* [6] and Deléger *et al.* [7, 8], with German and French as the target languages, respectively. Both studies report precision values of about 80%.

However, due to inconsistent evaluation strategies used in the literature, the influence of the used resources as well as language-specific system design and tuning decisions, it is hard to generalize from these results.
Terminology extraction in the biomedical field is tricky, as most terms are not merely single words (e.g. *'appendicitis'*) but multi-word expressions (MWEs), like *'Alzheimer's disease'* or *'acquired immunodeficiency syndrome'*. Approaches towards finding MWEs can be classified as either pattern-based, using e.g. manually created part-of-speech (POS) patterns, or statistically motivated, utilizing e.g. phrase alignment techniques. The former solution (used e.g. by Déjean *et al.* [6] or Bouamor *et al.* [9]) suffers from the need to supply POS patterns which are often hand-crafted and may become cumbersome to read and write as the pattern set keeps growing. Statistical approaches circumvent this dilemma and can use e.g. the translation probabilities of the single words of a MWE (treated as a bag of words) [10] or some kind of phrases. These can either be linguistically motivated, i.e. use POS information [11], or be purely statistical and derived from the translation model produced by a phrase-based SMT system [12], just like in our system.

4 Conclusion

We described a system using SMT and NER to generate new entries for multilingual biomedical terminologies. A terminology enriched this way can be used to improve the annotation of raw language data corpora.
A direct evaluation of terminology enrichment systems like ours is complicated for two reasons. First, a missing standard metric—some report only precision based on the number of correct translations produced by their system [7], while others issue F-scores based on the system's ability to reproduce a (sample) terminology [6]—and, second, the rather strong influence of the chosen terminology and corpora on a system's performance. The CLEF-ER 2013 Challenge will allow us to overcome this problem, by enabling extrinsic comparison based on the annotations provided by the different systems.

Acknowledgments

This work is funded by the European Commission's 7th Framework Programme for small or medium-scale focused research actions (STREP) from the Information Content Technologies Call FP7-ICT-2011-4.1, Challenge 4: Technologies for Digital Content and Languages, Grant No. 296410.

References

1. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29**(1) (2003) 19–51
2. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: MOSES: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association

for Computational Linguistics Companion Volume. Proceedings of the Interactive Poster and Demonstration Sessions, Prague, Czech Republic, June 2527, 2007. Association for Computational Linguistics (2007) 177–180
3. Hahn, U., Buyko, E., Landefeld, R., Mühlhausen, M., Poprat, M., Tomanek, K., Wermter, J.: An overview of JCoRe, the JULIE Lab UIMA component repository. In: Proceedings of the LREC'08 Workshop 'Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP', Marrakech, Morocco, 31 May 2008. Paris: European Language Resources Association (ELRA) (2008) 1–7
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. ACM SIGKDD Explorations **11**(1) (2009) 10–18
5. Tiedemann, J.: News from OPUS: A collection of multilingual parallel corpora with tools and interfaces. In: Recent Advances in Natural Language Processing. Volume V. Amsterdam, Philadelphia: John Benjamins (2009) 237–248
6. Déjean, H., Gaussier, E., Renders, J.M., Sadat, F.: Automatic processing of multilingual medical terminology: applications to thesaurus enrichment and cross-language information retrieval. Artificial Intelligence in Medicine **33**(2) (2005) 111–124
7. Deléger, L., Merkel, M., Zweigenbaum, P.: Enriching medical terminologies: an approach based on aligned corpora. In Hasman, A., Haux, R., van der Lei, J., De Clercq, E., Roger France, F.H., eds.: MIE 2006 – Proceedings of the 20th International Congress of the European Federation for Medical Informatics. Volume 124 of Studies in Health Technology and Informatics. Maastricht, The Netherlands, August 27-30, 2006. Amsterdam: IOS Press (2006) 747–752
8. Deléger, L., Merkel, M., Zweigenbaum, P.: Translating medical terminologies through word alignment in parallel text corpora. Journal of Biomedical Informatics **42**(4) (2009) 692 – 701
9. Bouamor, D., Semmar, N., Zweigenbaum, P.: Identifying bilingual multiword expressions for statistical machine translation. In: LREC 2012 – Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey, 21-27 May 2012. European Language Resources Association (ELRA) (2012) 674–679
10. Vintar, Špela: Bilingual term recognition revisited; the bag-of-equivalents term alignment approach and its evaluation. Terminology **16**(2) (2010) 141–158
11. Lefever, E., Macken, L., Hoste, V.: Language-independent bilingual terminology extraction from a multilingual parallel corpus. In: EACL 2009 – Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece, 30 March 3 April, 2009. Association for Computational Linguistics (2009) 496–504
12. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: NAACL'03 – Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Volume 1., Edmonton, Alberta, Canada, May 27 - June 1, 2003. Association for Computational Linguistics (2003) 48–54