Porting a Rule-based Assertion Classifier for Clinical Text from English to Swedish

Sumithra Velupillai[1], Maria Skeppstedt[1], Maria Kvist[1,2], Danielle Mowery[3],
Brian E. Chapman[4], Hercules Dalianis[1] and Wendy W. Chapman[4]

[1]Dept. of Computer and Systems Sciences (DSV), Stockholm University, Sweden
{sumithra,mariask,hercules}@dsv.su.se
[2]Dept. of clinical immunology and transfusion medicine, Karolinska University Hospital, Sweden maria.kvist@karolinska.se
[3]Dept. of Biomedical Informatics, University of Pittsburgh dlm31@pitt.edu
[4]Division of Biomedical Informatics, University of California San Diego
{brchapman,wwchapman}@ucsd.edu

Abstract. An existing rule-based assertion classifier is ported from English to Swedish: pyConTextSwe. Evaluation on Swedish clinical texts shows that the English lexical resources are useful, but that there are assertion cues not obtainable in existing resources. Iterative error correction of cue lexicons improves results for the ported classifier. Overall final results are 82% F-score on a development set and 74% on a test set.

1 Introduction

Assertion classification - negation and uncertainty - is a critical task in information extraction of disorder mentions in clinical texts. Although assertion classification for English has received considerable attention, assertion research for other languages is still in its early stages [1], [2], [3], [4]. Developing algorithms for a new language can be tedious and challenging since obtaining and annotating data takes time and resources. Utilizing existing systems and lexical resources developed for another language could reduce this burden. In our aim to study the feasibility of porting existing approaches between languages, we used one language pair as an example (English vs Swedish) for this study with three goals: 1) develop an assertion lexicon for NLP of Swedish clinical texts based on translations of existing English lexicons, 2) port an existing assertion algorithm from English to Swedish (pyConTextSwe), 3) provide an analysis of the issues involved in translating lexical cues from English to Swedish for this task. To address these goals, we evaluated the portability of a simple, existing assertion classification system, pyConTextNLP [5], on Swedish clinical texts.

The 4th International Louhi Workshop on Health Document Text Mining and Information Analysis (Louhi 2013), edited by Hanna Suominen

2 Background

Stating whether a disorder is negated or affirmed is not always a binary choice, but can be modeled as a continuum ranging from definitely positive to negative [5], [3]. In clinical texts, care providers often make use of epistemic modality and linguistic hedging to indicate a level of belief or confidence about whether a disorder exists. For instance, in *Patient most likely has pneumonia*, the care provider asserts that pneumonia is a probable diagnosis with certainty expressed as a high probability.

Negation and uncertainty identification have been the focus of several shared tasks including the 2010 i2B2/VA Challenge [6], CoNLL-2010 [7], and BioNLP 2009 [8] in clinical, biomedical, and biological texts, respectively. For negation, many NLP tools achieve high performance for identifying negation of disorders using cue lexicons and heuristics, including NegEx [9], NegFinder [10], and NegExpander [11]. For uncertainty, rule based and machine learning approaches have shown variable performance for predicting uncertainty as a binary choice or as a continuum, e.g. [6], [5], [4]. Lexical knowledge generated from an existing English approach, NegEx, has been utilized for other languages such as Swedish (F-score 78%) [1] and French (F-score 87%) [2] with performance differences driven mainly by precision at 75% and 89%, respectively.

The existing assertion classifier used in this study is pyConTextNLP, a Python implementation of the ConText algorithm that includes a generalization of the concepts of targets and cues to include any type of relationship specified by the user. pyConTextNLP has previously been used for identifying and characterizing pulmonary embolism findings in CT pulmonary angiography reports [5].

3 Methods

Two subsets of a corpus in Swedish annotated for uncertainty on a diagnostic statement level was used for this study, see Table 1. This corpus contains assessment entries from an emergency department in Karolinska University Hospital.[1]

Porting pyConTextNLP to Swedish involved four steps: (1) translation and correction of lexical cues to Swedish, (2) alteration of the classifier to process non-English text, (3) creation of a wrapper application: pyConTextSwe and (4) iterative modification of pyConTextSwe and lexicons through error analysis on a development set. Results were evaluated with precision, recall and F-score.

(1) We employed an additive approach for creating a number of Swedish cue lexicons, starting with the lexical cue set that was simplest to generate and adding new sets to the evaluation of the assertion classifier. The baseline lexicon was developed by applying Google Translate to the English cue lexicon used in pyConTextNLP[2] (GT). One physician and one computational linguist independently corrected the GT lexicon then discussed their corrections in a group with

[1] Ethical approval number 2009/1742-31/5

[2] http://code.google.com/p/negex/wiki/pyConTextNLP

Table 1. Dataset: number of instances, development (devel) and test (test) set, with original annotation classes and mapped output values. Mapping is based on previous studies [3]

Gold	*devel*	*test*	**Mapped**	*devel*	*test*
Certainly Positive	82	148	Definite existence	82	148
Probably Positive	50	49	Probable existence	89	93
Possibly Positive	39	36			
Possibly Negative	0	9			
Probably Negative	26	24	Probable negated existence	26	24
Certainly Negative	44	27	Negated existence	44	27
Total	241	292		241	292

Table 2. Results: iterations with different cue lexicons, all four classes. GT = Google Translate of existing English cue lexicon. GC = Google Translate Corrected. E = Extra cues from other versions of PycontextNLP and Swedish guidelines. EA = Additions from error analysis. Cues = cues in lexicon (number of cues actually found + combined cues found through precedence rules).

Annotation class	Lexical cue set	Cues	Precision (%)	Recall (%)	F-score
definite existence	GT	3 (0)	40.74	**93.90**	56.83
	GC	5 (0)	53.52	92.68	67.86
	GC+E	11 (1)	67.59	89.02	76.84
	GC+E+EA	11 (1)	84.88	89.02	**86.90**
	Test: GC+E+EA	11 (1)	84.46	84.46	84.46
probable existence	GT	49 (5+2)	73.33	24.72	36.97
	GC	85 (6)	89.29	28.09	42.74
	GC+E	136 (18+1)	82.26	57.30	67.55
	GC+E+EA	153 (24+5)	**89.74**	**78.65**	**83.83**
	Test: GC+E+EA	153 (22+5)	70.41	74.19	72.25
probable negated existence	GT	66 (3)	44.44	15.38	22.86
	GC	83 (2+3)	**75.00**	23.08	35.29
	GC+E	88 (3+6)	40.62	50.00	44.83
	GC+E+EA	94 (7+8)	59.38	**73.08**	**65.52**
	Test: GC+E+EA	94 (2+2)	46.67	29.17	35.90
definite negated existence	GT	56 (2)	46.15	13.64	21.05
	GC	65 (6)	63.49	**90.91**	74.77
	GC+E	75 (6+1)	64.10	56.82	60.24
	GC+E+EA	77 (9)	**77.78**	79.55	**78.65**
	Test: GC+E+EA	77 (5)	48.39	55.56	51.72

a computational linguistics student and a native English speaking medical informaticist until there was consensus about the best corrections (GC). This step also involved adding alternative translations for some cues. We enriched the lexicon with translated lexical cues from previous versions of PyContextNLP, cues from Swedish annotation guidelines, and with cues generated manually by the physician (E). (2) For porting, we made a number of changes to the pyConTextNLP package including providing full support for Unicode texts, and providing improved modularity for customization, e.g. with defining lexical rules and sentence splitting rules. (3) pyConTextSwe maps the output of pyConTextNLP to values represented in the manual annotations (Table 1). Further, it contains precedence rules for handling potentially conflicting cues, e.g. when two cues are found for the same diagnostic statement, e.g. *ej sannolik diabetes* (not likely diabetes), where *ej* = negated existence and *sannolik* = probable existence, generating

Table 3. Results: existence yes/no and uncertainty yes/no. GC = Google Translate Corrected. E = Extra cues from other versions of PycontextNLP and Swedish guidelines. EA = Additions from error analysis.

Annotation class	Lexical cue set	Precision (%)	Recall (%)	F-score
existence_yes	GC+E+EA	96.95	92.98	94.93
	Test: GC+E+EA	94.72	96.68	95.69
existence_no	GC+E+EA	84.42	92.86	88.44
	Test: GC+E+EA	82.61	74.51	78.35
uncertainty_yes	GC+E+EA	86.36	82.61	84.44
	Test: GC+E+EA	72.57	70.09	71.30
uncertainty_no	GC+E+EA	84.73	88.10	86.38
	Test: GC+E+EA	80.45	82.29	81.36

the output value probable negated existence. (4) Finally, we added cues found through error analysis (EA) on the output of running pyConTextSwe on the development set (Table 1). For final evaluation, we used the accumulated cue lexicon on a separate test set, measuring performance on unseen instances.

4 Results

Overall results for each output value is best when using the accumulated lexicon (GT+GC+E+EA), Table 2. However, GC yields best precision results for probable negated existence and best recall for definite negation (75.00 and 90.91 respectively). For definite existence, recall results are high already when using GT (93.9), since this is the default value. Results on the test set are lower (81.74 vs 73.97, micro-averaged overall F-score).

Results for the binary classifications existence/uncertainty yes/no, also show a drop in performance between the development set and test set for uncertainty yes and existence no (F-score 84.44 to 71.30 and 88.44 to 78.35 respectively), Table 3. However, this drop is smaller than for the multi-class classification task, especially compared to the drop for probable and definite negated existence.

Only a small proportion of the cues in the accumulated lexicon are found in the data. In Figure 1 we see that only 29 cues account for a recall result of 79% in the development data, for the output value probable existence, and even fewer for probable negated and definite negated.

5 Discussion

We present a study on porting an assertion classifier from English to Swedish using different cue lexicons. Overall F-score results are 82% on a development set and 74% on a test set using a lexicon enriched with cues found through error analysis. Testing pyConTextSwe on unseen data yielded lower results on all classes, but two proved to be more problematic, i.e. probable negated existence and definite negated existence, indicating that further analysis and definitions are needed for this distinction.

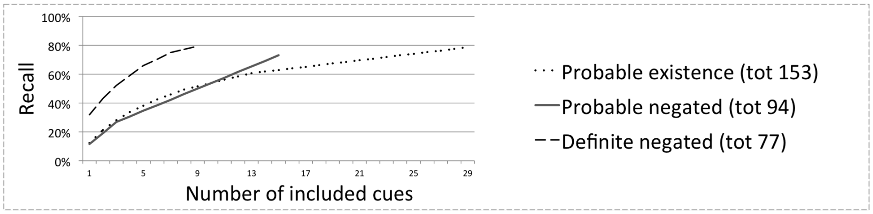

Fig. 1. Number of cues, ordered by frequency of the cue in the data (x-axis) accounting for recall results (y-axis) for the output values probable existence, probable negated existence and definite negated.

Previous results for assertion classification are difficult to compare for several reasons; in the definition of the task itself, representation model of negation and uncertainty, and data used for the tasks. However, the general trends are comparable. For instance, for negation detection, our results are in line with those presented for Swedish [1] and for English [9], but lower than those presented for French [2]

Overall assertion classification results in the 2011 i2b2 challenge [6] are 93.6 F-score, which is higher compared to our 73.97, but the definition of assertions differs greatly. When compared to a machine-learning based approach for a similar task in Swedish, results are similar for some classes but lower for others (e.g. 56.4 vs 35.9 F-score for probable negation) [4].

When adding cues from an error analysis, only two extra cues for negation were added, whereas more were added for other classes, Table 2. An error analysis was also performed on the test set. A total of 42 potential new cues were found, three for definite existence, six for definite negation, 15 for probable negation, 18 for probable existence. This shows that the lexicons were not exhaustive enough to capture all assertion expressions in the test data.

A large proportion (74%) of these cues were versions of existing cues, e.g. inflections (13%) or cue expressions partially contained in existing cues, but with an extra or another modifier (51%). For instance, the translation of *suspicion* was classified as probable existence, but in the test data, this cue occurs in the context *low suspicion*, thus *probable negated.* Likewise, the translation of *speaks for* is in the cue lexicon, but it does not match *speaks mostly for* from the test data. Translated cues account for two thirds of the found cues for definite negation. Because of larger variability in expressing uncertainty (Figure 1), only one fourth of the found cues originate from translations. This will be further examined in studies on cue coverage in larger clinical corpora.

Porting pyConTextNLP to pyConTextSwe worked well, but emphasis must be put in developing useful lexicons, especially for uncertainty cues. Translation of cues needs to be complemented with additional methods for lexicon expansion. Since a large proportion of cues found in the data were versions of existing cues, it is not sensible to manually search for speculation cues. We have the intention to apply methods for automatic generation of inflected cues (or for lemmatization), and methods for generation of word collocations. We also plan to apply methods for automatic synonym extraction.

There are a number of additional lexical resources that could be utilized for assertion classification, such as SNOMED CT and cues in the BioScope corpus [7]. As the next step to further improve pyConTextSwe, we have translated and classified assertion cues from the BioScope corpus and have also extracted cues from another Swedish annotated resource. These cues, along with cues obtained from collocation and synonym extraction, will be included in the next version of pyConTextSwe. As a final step in our study of portability between Swedish and English, we have translated the obtained Swedish cues back to English and will close the circle by an evaluation on English clinical text, hopefully adding useful cues to the original English assertion classifier. Furthermore, we are involved in ongoing studies for comparing other language pairs and studying cue distributions in different clinical corpora and languages.

Acknowledgements
We would like to thank the Stockholm University Academic Initiative (Interlock project) for their funding support, and the anonymous reviewers for their comments.

References

1. Skeppstedt: Negation detection in Swedish clinical text: An adaption of NegEx to Swedish. Journal of biomedical semantics **2**(Suppl 3) (2011) S3
2. Deléger, L., Grouin, C.: Detecting negation of medical problems in French clinical notes. In: Proc. 2nd ACM SIGHIT, Intl. health informatics, ACM (2012) 697–702
3. Velupillai, S., Dalianis, H., Kvist, M.: Factuality Levels of Diagnoses in Swedish Clinical Text. In: Proc. 23rd MIE, Oslo, IOS Press (August 2011) 559 – 563
4. Velupillai, S.: Automatic Classification of Factuality Levels – A Case Study on Swedish Diagnoses and the Impact of Local Context. In: Proc. 4th LBM 2011, Singapore (December 2011)
5. Chapman, B., Lee, S., Kang, H., Chapman, W.: Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. J Biomed Inform **44**(5) (2011) 728–737
6. Uzuner, O., South, B., Shen, S., DuVall, S.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. JAMIA **18**(5) (2011) 552–556
7. Farkas, R., Vincze, V., Móra, G., Csirik, J., Szarvas, G.: The conll-2010 shared task: Learning to detect hedges and their scope in natural language text. In: Proc. 14th CoNLL, Uppsala, Sweden, ACL (July 2010) 1–12
8. Kim, J.D., Ohta, T., Pyysalo, S., Kano, Y., Tsujii, J.: Overview of BioNLP'09 shared task on event extraction. In: BioNLP '09, PA, USA, ACL (2009) 1–9
9. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform (2001) 34–301
10. Mutalik, P., Deshpande, A., Nadkarni, P.: Use of general-purpose negation detection to augment concept indexing of medical documents a quantitative study using the umls. JAMIA **8**(6) (2001) 598–609
11. Aronow, D., Feng, F., Croft, W.B.: Ad-hoc classification of radiology reports. JAMIA (1999) 393–411