CREATING A BILINGUAL PSYCHOLOGY LEXICON FOR CROSS LINGUAL QUESTION ANSWERING, A PILOT STUDY

Andrea Andrenucci
Department of Computer and System Sciences, Stockholm University/ Royal Institute of Technology,
Forum 100, SE-16440, Kista, Sweden
andrea@dsv.su.se

Keywords: Internet services, Natural Language Interfaces, Data Mining, Cross Lingual Question Answering

Abstract: This paper introduces a pilot study aimed at investigating the extraction of word relations from a sample of a medical parallel corpus in the field of Psychology. Word relations are extracted in order to create a bilingual lexicon for cross lingual question answering between Swedish and English. Four different variants of the sample corpus were utilized: word inflections with and without POS tagging, lemmas with and without POS tagging. The purpose of the study was to analyze the quality of the word relations obtained from the different versions of the corpus and to understand which version of the corpus was more suitable for extracting a bilingual lexicon in the field of psychology. The word alignments were evaluated with the help of reference data (gold standards), which were constructed before the word alignment process.

1. INTRODUCTION

Users of medical portals in general, regardless of their background, value the possibility of formulating their information needs in their own native language (Andrenucci, 2006). In Question Answering this is possible with the help of Machine Translation (MT), which converts user questions into the language of the texts from where the answers are extracted. This paradigm is called Cross Language Question Answering, CLQA (Aunino, Kuuskoski and Makkonen, 2004).

The Web4health medical portal (http://web4health.info) supports cross language question answering (CLQA). User questions are translated into English with the help of Systran's MT system (http://www.systransoft.com) and are then used to retrieve answers from the knowledge base of the portal. One problem with the existing implementation is that Systran implements medical lexicons which are not tailored to the specific domain of the portal, i.e. psychology and psychotherapy. The aim of the project presented in this paper is to produce a bilingual lexicon for Swedish and English that overcomes this gap. In order to achieve this goal we have investigated in a pilot study the possibility of automatically extracting word relations from a parallel corpus, which is a sample of Web4health's knowledge base. The sample corpus was extracted in two versions, one version consisting of words in their inflected forms and another version consisting of word lemmas. For both versions we also provided a variant annotated with part of speech (POS) tagging and a variant without POS tagging. The purpose of the study was to analyze the quality of the word relations obtained from the different versions of the corpus and to understand which version of the corpus was more suitable for extracting a bilingual lexicon. The texts were aligned at the paragraph, sentence and word level with the Uplug toolkit (see section 3), a collection of tools for processing parallel corpora, developed by Jörg Tiedemann (2003a). Uplug utilizes both statistical and linguistic information in the alignment process. The alignments were evaluated at the word level with the help of reference data (gold standard), which were constructed before the word alignment process (see section 4.5).

The paper is structured as follows: section two describes related research in the field of cross lingual question answering. Section three summarizes the knowledge base and the Uplug toolkit. Section four and five describe the pilot study

and its quantitative results. The paper is concluded with a discussion of the results (section six) and the paper conclusions (section seven).

2. RELATED RESEARCH

Several projects have focused on developing lexical resources for specific domains. For example Weijnitz et al. (2004) describes the implementation of a Swedish-English lexicon for the agricultural domain, which was then utilized to compare translations from two different MT systems: a system based on statistical tools such as the ISI ReWrite Decoder (Germann, 2003) and a rule based MT system. Loukachevitch and Dobrov (2004) developed a Russian-English thesaurus for the socio-political domain as a resource for automatic text processing and information retrieval. The thesaurus is based on definitions of taxonomic and ontological dependence relations between domain specific concepts.

For what concerns medicine, one of the most utilized lexical resources in QA and Information Retrieval is the Unified Medical Language System (UMLS, Lindberg, Humphreys, and McGray, 1993). It contains different knowledge resources such as lexicons and thesauri, and represents medical concepts with the help of semantic networks.

When it comes to the extraction of domain-specific bilingual dictionaries through word alignment (with Swedish as source or target language), previous research has mainly focused on comparing the quality of results with and without POS tagging (Nyström et al., 2006, Tiedemann, 2003a) and with shallow syntactic parsing (Tiedemann, 2003b). Lemmatized versions of the corpora were not included in the evaluations. Since the utilization of stemming in Swedish improves precision and recall in information retrieval (Carlberger et al., 2001), and word alignment can be viewed as a retrieval problem (Ahrenberg et al., 2000), we have included lemmatization in our evaluation.

3. THE KNOWLEDGE BASE AND THE UPLUG TOOLKIT

The Web4health medical portal (http://web4health.info) is well established among the medical portals on the Web. It is Yahoo-listed and it was developed within a EU-financed project called KOM 2002, whose goal is to provide multilingual medical information to improve the mental health of European citizens. Psychiatrists and psychotherapists from five different European countries (Italy, Sweden, Holland, Greece and Germany) use the portal to jointly develop a set of semantically classified Web pages that answer questions in matters of psychological and psychotherapeutic advice. Users consult the knowledge base submitting questions in natural language, which are then matched against pre-stored FAQ-files (Frequently Asked Questions) consisting of question/answer pairs, where the question part has a template created to match many different variations of the same question (Template-Based Question Answering, Sneiders 2002).

The Uplug toolkit (Tiedemann, 2003a) is a collection of tools for processing parallel corpora. Its main functionality consists of sentence and word alignments of bilingual texts. The main idea behind Uplug's alignment process is to utilize both linguistic and statistical information in order to extract word relations. Each individual piece of information is called a clue, $C_i(s,t)$, and is defined as a probability that indicates an association between two sets of words s and t in parallel texts. Formally it is defined as a weighted association A between s and t, where w_i is used to weight and normalize the score of A_i:

$$C_i(s,t) = P(a_i) = w_i A_i(s,t) \qquad (1)$$

All clues are then combined in an overall measure, which is defined as the disjunction of all indications:

$$C_{all}(s,t) = P(a_{all}) = P(a_1 \cup a_2 \cup ... \cup a_n) \qquad (2)$$

Clues are not mutually exclusive. The addition rule for probabilities generates the following formula for a disjunction of two clues:

$$P(a_1 \cup a_2) = P(a_1) + P(a_2) - P(a_1 \cap a_2) \qquad (3)$$

Two main types of clues are considered: basic (static) clues, whose value is constant for a pair of lexical items and dynamic clues, whose values are learned dynamically during the alignment process. Basic clues include co-occurrence coefficients (the Dice coefficient, Tiedemann 1999), string similarity coefficients (the longest common subsequence ratio, Melamed, 1995) and GIZA++ clues (Och and Ney, 2003), based on IBM models (Brown et al., 1993) and Hidden Markov Model. Dynamic clues include patterns of POS labels, phrase types and word positions. The system aligns first sentences and words with the basic clues and then utilizes the aligned links as training data in order to learn new dynamic clues and improve the quality of the

alignments. For instance, examining POS tags in source and target language, it is possible to estimate the probabilities of translation relations between words that belong to certain word classes.

A huge advantage of the Uplug tool is that it supports the dynamic construction of alignments with multi word units (MWUs), i.e. noun phrases, idiomatic expressions and other phrasal constructions that should not be split up in the alignment process (Tiedemann, 2003b, p. 18).

4. IMPLEMENTATION OF THE PILOT STUDY

This section describes how the pilot study was conducted. It introduces some linguistic characteristics of Swedish in comparison to English (section 4.1) and then outlines how the sample corpora were selected (section 4.2), prepared for the alignment process (section 4.3), plus how the results were evaluated (section 4.4).

4.1 The Swedish Language in brief

Swedish is an inflective language that belongs to the Germanic branch of Indo-European languages. It has a more complex morphology than English. Gender, definiteness and plurality are suffixed in nouns and adjectives. Adjectives and articles agree with the head noun in terms of gender, definiteness and number. Genitive forms are formed by the suffix "s". Nouns have two genders: gender uter ("en"-words) and gender neuter ("ett"-words). Similarly to English, nouns can be written with or without articles and sentences implement the subject-verb-object order. Homographs and compound words are very common in Swedish. Compounds are often constructed with an extra consonant or vowel (called *fogemorphemes*) that joins the constituents of the compounds: e.g. "koncentrationssvårigheter" ("attention problems") is composed by putting together the words "koncentration" (attention) and "svårigheter" (problems), with the fogemorpheme "s" to bind them. Swedish has also particle verbs, i.e. compound verbs where one of the components is a particle. Particle verbs can either be tightly compounded, i.e. with the particle embedded in the verb as a prefix, e.g. "påminna" (to remind), or loosely compounded, i.e. with the particle coming after the verb, e.g. "tycka om" (to like).

4.2 The Corpus Selection

The parallel corpus utilized in this pilot study includes a randomly selected set of FAQs, i.e. question/answer pairs, in the source language (Swedish) and the target language (English). The Swedish corpus consists of circa 12800 tokens and the English counterpart of circa 13000 tokens.

Prior to utilizing the randomly chosen texts, we scanned and proofread the material and, when necessary, corrected it to ensure its completeness and correctness. This was a difficult and time consuming task, since the documents in the repository are often translated freely and the structure of the texts tend also to differ, with sentences or phrases that are available in one language only.

4.3 Annotating the Corpora

Prior to starting the alignment process, some preliminary work was needed in order to prepare the corpora. Since the FAQ documents are annotated with HTML tags, the texts had first to be cleaned up by the existing tags and then converted into plain text. The Uplug toolkit was then used for encoding the texts with ISO88591 for Latin1 (which includes Swedish and English) and annotating them with XML Corpus Encoding Standard (XCES) (Ide and Priest-Dorman, 2000). Sentence splitting and tokenization were included in this step. The sentences and words were marked with an ID-number.

A version of the bilingual corpus was lemmatized with the CST Lemmatizer (Jongejan and Haltrup, 2005) which is a trainable, rule-based tool that works with languages that utilize inflectional suffixes, such as Swedish and English.

The Trigrams'n Tags tagger (Brants, 2000) was utilized to annotate the POS-tagged versions of the corpora. TnT was chosen since it is the tagger that has the highest overall accuracy among data-driven taggers and succeeded best in the annotation of both known and unknown words in Swedish (Megyesi, 2000).

The tagger was trained on Swedish (Megyesi 2002) using the StockholmUmeå Corpus (SUC, 1997), and utilized for the labels the PAROLE annotation scheme (Ejerhed and Ridings, 1995), a tagset that include part-of-speech and morphological features such as gender and number of the words. The Penn Treebank corpus and its tagset (Marcus, Santorini och Marcinkiewicz, 1993), which also encodes morphological information such as number, were utilized for the English language.

4.4 Extraction of Word Relations

After aligning the different versions of the corpus at the sentence level, capital letters were converted to non-capital letters in order to improve precision of the word-level alignment. Once the word alignment was finished, a table, with word-pair frequencies sorted in descending order, was constructed for each corpus version in order to see which alignments occurred more often. These frequency tables were later utilized for analyzing the evaluation results (see section 5 and 6).

4.5 Evaluation method and the Gold Standards

Two main evaluation techniques are utilized when it comes to evaluating word alignment (Ahrenberg et al., 2000): automatic evaluation with a reference alignment (Gold Standard) or manual evaluation by experts. Automatic evaluation was preferred since reference alignments can be re-utilized and it is possible to control the process of selecting the reference data, focusing for instance on certain word types or words from certain frequency ranges (Merkel, 1999).

Two gold standards, consisting of 130 items each, were developed for the evaluation. They were aligned manually according to detailed guidelines (Merkel, 1999). The first GS was compiled by randomly selecting word samples from the parallel corpus. The word samples were limited to content units (phrases and content words, i.e. words with a full meaning of their own), since the purpose of our research is to extract a bilingual lexicon that is specific for the psychological domain. We applied a frequency balanced approach, i.e. we grouped entries according to the following frequency ranges: 10 entries with frequency above 10, 30 entries with frequency 7-9, 30 with frequency 5-6, 30 with frequency 3-4 and 30 with frequency 1-2.

Similarly the second GS was compiled following the same approach, but utilizing as information source the sets of all user queries submitted to Web4health portal. Both GS included links of type "regular" (standard), "fuzzy" (somehow semantically overlapping but with different POS or different degrees of specification) and "null" (omissions). Complex MWU links were also included.

As stated of Ahrenberg et al. (2000), word alignment can be viewed as a retrieval problem. For this reason, when evaluating the quality of the alignments, it is appropriate to apply measures from the field of information retrieval such as precision and recall. By precision it is meant the ratio of correctly aligned items in proportion to the number of aligned items and by recall the ratio of correctly aligned items in proportion to the total number of correct items (reference data). However a problem of these measures is that they do not handle *partially correct links*, i.e. links that have at least one correct word on source and target side, since links are either considered as entirely right or entirely wrong. This approach works well when it comes to evaluating single word alignments, but is too coarse for the evaluation of MWUs, which often imply partially correct results (Tiedemann, 2003b, p. 26).

In order to overcome this deficiency we chose to apply refined metrics of precision and recall (Tiedemann, 2003b, p. 68) that measure the *degree of correctness* of the proposed links. They calculate a partiality value Q that is proportional to the number of words that are in common between the proposed alignments and the reference data:

$$Q_x^{precision} = \frac{\left|aligned_{src}^x \cap correct_{src}^x\right| + \left|aligned_{trg}^x \cap correct_{trg}^x\right|}{\left|aligned_{src}^x\right| + \left|aligned_{trg}^x\right|} \quad (4)$$

$$Q_x^{recall} = \frac{\left|aligned_{src}^x \cap correct_{src}^x\right| + \left|aligned_{trg}^x \cap correct_{trg}^x\right|}{\left|correct_{src}^x\right| + \left|correct_{trg}^x\right|} \quad (5)$$

$aligned_{src}^x$ is the set of source language words and $aligned_{trg}^x$ the set of target language words in link proposals for a reference link x in the GS. $correct_{src}^x$ and $correct_{trg}^x$ define the sets of source and target words of reference link x. Precision (P) and recall (R) are then defined with the help of Q:

$$R = \sum_{x=1}^{X} \frac{Q_x^{recall}}{\left|correct\right|} \quad P = \sum_{x=1}^{X} \frac{Q_x^{precision}}{\left|aligned\right|} \quad (6)$$

$\left|aligned\right|$ is the total number of correct, incorrect and partially correct links in relation to the GS and $\left|correct\right|$ represents the size of the GS.

These metrics handle also partially correct links in a more fine grained way, unlike other coarser approaches (e.g. the PLUG metrics, Ahrenberg et al., 2000) that penalize partially correct links with a constant value without considering the degree of correctness of the links.

Table 1 below shows some examples taken from an evaluation protocol of the word alignments produced for the corpus version with POS tagged word forms and with respect to the first GS. The precision, recall and F-score results are calculated with the aforementioned metrics.

Table 1: Examples from an evaluation protocol.

Type	ID	Source	Target
correct	SL4.16	muskelsvaghet	muscular weakness
correct	SL5.12	förvränger	distorts
partial 2(3)	SL22.5	missbruka	to using (using)
GS size:	130	regular: 125	fuzzy: 3 null: 2
Correct:	74	regular: 73	fuzzy: 1 null: 0
Partially correct:	49	regular: 48	fuzzy: 1 null: 0
Incorrect:	7	regular: 4	fuzzy: 1 null: 2
Missing links:	0	regular: 0	fuzzy: 0 null: 0
Precision:	76.52%		
Recall:	89.10%		
F-score:	82.34%		

5. QUANTITATIVE RESULTS

Our study produced the quantitative results that are shown in table 2 and table 3 below. Table 2 shows the results of precision, recall and F-score for all the corpus versions and in relation to the two GS. Table 3 presents the number of correct, partially correct and erroneous links calculated also in relation to both GS. In the next section we discuss the results with the help of the data from the frequency tables and elicit the differences between each version of the corpora.

Table 2: Precision, Recall and F-score results.

Corpus based GS (first GS)			
Type of Corpus	Precision	Recall	F-score
Lemmas no POS	77.38%	88.69%	82.65%
Lemmas with POS	78.08%	90.72%	83.93%
Word forms no POS	72.09%	87.58%	79.08%
Word forms with POS	76.52%	89.10%	82.34%
Query based GS (second GS)			
Lemmas no POS	68.95%	86.78%	76.84%
Lemmas with POS	69.52%	86.22%	76.97%
Word forms no POS	67.56%	87.24%	76.15%
Word forms with POS	71.21%	87.49%	78.51%

The results in table 2 do not present striking differences between the corpora, however the POS tagged lemmas achieved slightly better results for the corpus based GS, which consisted of a larger number of single word units, and the POS tagged word forms obtained slightly better values with the query based GS, where a larger number of MWUs was included. Word forms without POS tagging achieved the lowest results with both GS. For what concerns the results in table 3 it is interesting to point out that word forms with POS had the highest number of correct links both with the corpus GS (together with the lemmas without POS) and with the query based GS.

Table 3: Number of Correct, Partially Correct and Incorrect Links out of the first and second GS.

Corpus based GS, size=130 links			
Type of Corpus	Correct	Partial	Incorrect
Lemmas no POS	74	49	7
Lemmas with POS	71	54	5
Word forms no POS	64	58	8
Word forms with POS	74	49	7
Query based GS, size=130 links			
Lemmas no POS	59	63	8
Lemmas with POS	59	63	8
Word forms no POS	56	66	8
Word forms with POS	67	54	9

6. DISCUSSION AND ANALYSIS

As a complement to the statistical data presented in table 2 and 3 we analyzed the frequency tables extracted from the alignments and compared the results, trying to elicit the similarities and the differences among the different versions of the corpus. We discuss our analysis with the help of some examples of the word relations that are presented in tables 4, 5, 6 and 7.

Lemmas with POS Tags VS Lemmas without POS Tags

The statistical results for Lemmas with and without POS were very similar with both GS. However the examination of the frequency tables clarified some points of difference. POS tagged lemmas were more precise when aligning compound words (e.g. “tvångsstörning - obsessive compulsive disorder” VS “tvångsstörning - obsessive compulsive”), in particular those with low frequency rate (1 or 2) in the texts. The POS tagged alignments had also fewer additions, i.e. words that are occurring in the alignments but that are not present in the reference links (“vätskedrivande - diuretic” VS “vätskedrivande som - diuretic”). Table 4 presents some other examples of those differences. Alignments consisting of words with similar strings and lengths achieved good quality in both cases (“söka - seek”, “eliminera - eliminate”, “effektivt -

effectively"), as well as words with high co-occurrence coefficients but dissimilar strings ("försöka - try", "mättnad - satisfaction").

However the POS tagging proved to be useful in aligning words consisting of dissimilar strings and with low co-occurrence frequency, but sharing the same POS (e.g. two adjectives: "betydande - significant" VS an adjective and a noun: "betydande - beginner").

Table 4: Alignment examples of lemmas with and without POS.

Lemmas no POS	Lemmas with POS
tvångsstörning - obsessive compulsive	tvångsstörning - obsessive compulsive disorder
trotssyndrom - oppositional defiant	trotssyndrom - oppositional defiant disorder
viktreglering – weight	viktreglering - weight control
vätskedrivande som - diuretic	vätskedrivande - diuretic
skräpmat - junk	skräpmat - junk food
betydande - beginner	betydande - significant
kontorsstol för – desk chair	kontorsstol - desk chair
här sjukdom - illness	sjukdom - illness

Inflected words with POS tags VS Inflected words without POS tags

As shown in table 2 and 3, alignments of inflected words with POS obtained, in comparison to inflected words without POS, better precision, recall and F-score results as well as a higher number of correct links. This confirms the results obtained by Nyström et al. (2006) and Tiedemann (2003b).

The POS tagged version produced more precise results both for MWUs and SWUs (see table 5). For what concerns single word units the morphological information of the POS tag was helpful for aligning words sharing the same definiteness ("förmågan - the ability" VS "förmågan - ability", "sjukdomen - the disease" VS "sjukdomen - diseases") or POS (e.g. two adjectives: "felaktiga - inappropriate" instead of a noun and an adjective: "antaganden - inappropriate").

In MWUs it was evident the role of POS for extracting links containing nouns with the same number ("barndomsupplevelser - childhood experiences" VS "barndomsupplevelser - childhood"). POS helped also to disambiguate the gender of Swedish adjectives in noun phrases, including them in the alignment when they agreed with the head noun and their inclusion was necessary to build a conceptual unit ("dåligt uppförande - misbehaviour" VS "uppförande - misbehaviour", where "dåligt" means "bad" and "uppförande" means "behaviour").

The POS based word relations had also better alignments among phrasal verbs that consisted of a verb and a particle in Swedish and a verb in English ("tänka ut - decide" VS "tänka - decide"; "klara av - handle" VS "klara - handle"). They even provided better alignments of verbs in passive forms ("uppfattas - are recognized" VS "uppfattas - regognized") and more fine grained links division ("möta - cope", "strategier - strategies" VS "strategier möta - strategies cope").

Table 5: Alignment examples of inflected words with and without POS.

Word inflections no POS	Word inflections with POS
aptitlöshet - appetite	aptitlöshet - loss of appetite
uppförande - misbehaviour	dåligt uppförande - misbehaviour
tänka - decide ut vettiga - on sensible	tänka ut - decide vettiga - sensible
barndomsupplevelser - childhood	barndomsupplevelser - childhood experiences
diet - intake	diet - food intake
förestaller - function innate	forestaller - picture
ångestdagbok panik - panic diary	ångestdagbok - panic diary
uppfattas - recognized	uppfattas - are recognized
tanker överdrivet - preoccupied	tänker överdrivet - are preoccupied
kostrådgivning - counselling	kostrådgivning - diet counselling
förmågan - ability	förmågan - the ability

Inflected words with POS tags VS Lemmas with POS tags

The statistical results in table 2 show that POS tagged lemmas produced slightly more precise alignments with the corpus based GS, while POS tagged word inflections were slightly more precise with the GS based on user queries. Inflected words had also a higher number of correct links with both GS (see table 3). The analysis of the alignment tables confirmed these results. POS tagged lemmas had more precise alignments of single unit words sharing the same lemma, since word inflections in the texts were converted into their base forms and were treated as the same word. This increased their co-occurrence frequency (see section 3), one of the basic clues of Uplug. For instance "trotsig" and "oppositional" co-occurred, in their lemmatized form, six times in the bitext, while as inflected forms they appeared two times as "trotsig - oppositional",

twice as “trotsiga - oppositional” and twice as “trotsigt - oppositional”.

Table 5: Alignment examples of inflected words and lemmas with POS.

Word inflections with POS	Lemmas with POS
andningsstörningar - breathing abnormalities	andningsstörning - abnormality
handla mat - shopping	handla - shopping
bufféserveringar - buffet services	bufféservering - progress buffet service
skönhetsidealer - beauty ideals	skönhetsideal - ideal
stämbanden - vocal cords	stämband - cord
vanligt förekommande - common	vanlig - common
barndomsupplevelser- childhood experiences	barndomsupplevelse - childhood experience contact
tanker ut - decides	kunna tänka ut - handla be decide
näringsbehoven - nutritional needs	näringsbehov - need
dåligt uppförande - misbehaviour	uppförande - misbehaviour

However the removal of number, definiteness and gender information in Swedish nouns and adjectives, obtained through lemmatization, determined a coarser POS tagging, affecting the dynamic clues (see section 3) and worsening the quality of the produced MWUs (see table 5). For instance removing the gender suffix in adjectives made it more difficult to individuate the nouns the adjectives referred to, causing less precise alignments in comparison to inflected forms (e.g. “uppförande - misbehaviour” instead of “dåligt uppförande - misbehaviour”). Furthermore wrongly lemmatized words caused erroneous POS tagging, which led to less accurate MWU alignments as well. For instance removing the suffix “t” from Swedish adverbs (e.g “vanligt”), as if they were adjectives referring to “ett-”words, made the tagger mark those words as adjectives instead of adverbs. This caused the omission of words that were necessary for the composition of MWUs (“vanlig - common” instead of “vanligt förekommande - common”).

Inflected words without POS tags VS Lemmas without POS tags

Lemmas without POS achieved better statistical results than inflected forms without POS in relation to both GS (see table 2 and 3). These results were also confirmed by analyzing the alignments produced with both forms. The corpus versions without POS tagging could not benefit from one important dynamic clue, the POS patterns. This implied that the other clues had a bigger influence on the clue alignment process. Two of those clues were the string similarity coefficient and co-occurrence coefficient. The inflected forms presented several erroneous alignments caused by string similarity of unrelated words and low co-occurrence of correct words, since inflections of the same word were considered by the system as different words. Lemmatization clumped inflected words to the same base word, increasing their occurrence frequency, and reducing their string length. This avoided some erroneous alignments that occurred in word inflections. For example the corpus with word inflections produced an alignment between the Swedish adjective “felaktigt” and the English gerund verb form “eating”; however the lemmatized form of “felaktigt”, “felaktig”, co-occurred often in the lemmatized corpus with the word “mistake” and was too dissimilar from the base form of “eating”, “eat”. This generated the link “felaktig - mistakte” instead of the erroneous “felaktigt - eating”.

There were no particular differences in the quality of alignments of proper nouns such as medicine names (Concerta - Concerta, Buspiron - Buspiron), since they were not subjected to inflections.

7. CONCLUSIONS

This paper has examined the extraction of word relations from a sample of a medical parallel corpus in order to create a bilingual lexicon for cross lingual question answering between Swedish and English. Four different variants of the sample corpus were created: word inflections with and without POS tagging, lemmas with and without POS tagging.

Inflected forms without POS tagging achieved the lowest results and it is not advisable to utilize them for the extraction of bilingual lexicon. POS tagging enhanced the quality of alignments of both SWUs and MWUs for both inflected forms and lemmas, especially of units with low frequency rate in the corpora or units consisting of dissimilar strings sharing the same POS.

POS tagged lemmas had slightly more precise alignments than POS tagged inflected words when it comes to SWUs. Lemmatization converted word inflections into their base forms, increasing their co-occurrence coefficients, since they were treated as the same word. However the information about gender, number and definiteness contained in the

suffixes of word inflections was crucial for the quality of alignment of MWUs. Considering that multi word terms were present in a larger number in the GS based on user queries and that the medical domain is characterized by MWUs, either unknown to generic lexicons or with meanings specific to this domain (Rinaldi et al, 2004), it is advisable to utilize corpora with POS tagged inflections as source for the extraction of bilingual lexicons for CLQA. Lemmatization should be applied on the frequency tables, after producing the word alignments, in order to group together words sharing the same base form in the source language or target language and facilitate the extraction of synonym lists in both languages.

As further work we intend to produce a follow-up study based on the whole content of the corpus as information source. We also intend to utilize larger gold standards (250 items per GS) and to provide statistical information (precision, recall and F-score) for each frequency range of the items in the GS.

REFERENCES

Ahrenberg, L., Merkel, M., Sågvall Hein, A., & Tiedemann, J., 2000. Evaluation of Word Alignment Systems. In *LREC'00, 2nd International Conference on Linguistic Resources and Evaluation*.

Andrenucci, A., 2006. Medical Information Portals: an Empirical Study of Personalized Search Mechanisms and Search Interfaces. In *ICEIS'06, 8th International Conference on Enterprise Information Systems*. INSTICC Press.

Aunino, L, Kuuskoski, R., & Makkonen, J., 2004. Cross-Language Question Answering at the University of Helsinky. In *CLEF' 04, Cross Language Evaluation Forum*.

Brants, T., 2000. TnT – A statistical Part-of-Speech Tagger. In *ANLP-2000, 6th Conference on Applied Natural Language Processing*.

Brown, P., Della Pietra, S., Della Pietra, V., & Mercer, R., 1993. The mathematics of statistical machine translation Parameter estimation [Electronic version]. *Computational Linguistics, 19,* 263-311

Carlberger, J., Dalianis, H., Hassel, M.,& Knutsson O., 2001. Improving Precision in Information Retrieval for Swedish using stemming. In *NoDaLiDa 2001, 13th Nordic Conference on Computational Linguistics*.

Ejerhed, E., & Ridings, D., 1995. *Parole and SUC*, http://spraakbanken.gu.se/parole/sgml2suc.html.

Germann, U., 2003. Greedy decoding for statistical machine translation in almost linear time. In *HLT-NAACL'03*. ACL press.

Ide, N., & Priest-Dorman, G., 2000. *Corpus encoding standard – document CES 1*. Technical report, Vassar College, LORIA/ CNRS. Vandoeuvre-les-Nancy, France.

Jongejan, B., & Haltrup, D., 2005. *The CST Lemmatiser*, Retrieved October 10, 2006, from Copenhagen University: http://cst.dk/download/cstlemma/current/doc/.

Lindberg, D., Humphreys, B., & McCray, A., 1993. The Unified Medical Language System [Electronic version]. *Methods of Information in Medicine, 32*, 281-291.

Loukachevitch, N., & Dobrov, B., 2004. Development of Bilingual Domain-Specific Ontology for Automatic Conceptual Indexing, In *LREC'04*.

Marcus, M., Santorini, B., & Marcinkiewicz, M., 1994. Building a large annotated corpus of English: The Penn Treebank [Electronic version]. *Computational Linguistics, 19*.

Megyesi, B., 2000. Comparing Data-Driven learning algorithms for PoS tagging of Swedish. In *NoDaLiDa2001*.

Megyesi, B., 2002. *DataDriven Syntactic Analysis Methods and Applications for Swedish*. PhD Thesis, Kungliga Tekniska Högskolan. Sweden.

Melamed, D., 1995. Automatic evaluation of uniform filter cascades for inducing N-best translation lexicons. In *3rd Workshop on Very Large Corpora*.

Merkel, M., 1999. *Annotation Style guide for the PLUG link annotator*. Technical Report, Linköping, University, Sweden.

Nyström, M., Merkel, M., Ahrenberg, L., et al., 2006. Creating a medical English-Swedish dictionary using interactive word alignment in BMC medical informatics and decision making [Electronic version]. *BMC Medical Informatics and Decision Making, 6*.

Och, F. J., & Ney, H., 2003. A Systematic Comparison of Various Statistical Alignment Models [Electronic version]. *Computational Linguistics, 29*.

Rinaldi, F., Dowdall, J., Schneider, G., & Persidis, A., 2004. Answering Questions in the Genomics Domain. In *ACL'04, Workshop on Question Answering in Restricted Domains*. ACL press.

Sneiders, E., 2002. *Automated Question Answering: Template-Based Approach*. PhD thesis, Royal Institute of Technology, Sweden.

SUC, 1997. *SUC 1.0 Stockholm Umeå Corpus, Version 1.0*. Umeå University and Stockholm University, Sweden.

Tiedemann, J., 1999. Word alignment – step by step. In *NODALIDA'99, the 12th Nordic Conference on Computational Linguistics*.

Tiedemann, J., 2003a. Combining Clues for Word Alignment. In *EACL'03, 10th Conference of the European Chapter of the ACL*. ACL press.

Tiedemann, J., 2003b. *Recycling translations. Extraction of lexical data from parallel corpora and their*

application in natural language processing. PhD thesis, Uppsala University, Sweden.

Weijnitz, P., Forsbom, E., Gustavii, E., Pettersson, E.,& Tiedemann, J., 2004. MT goes farming: Comparing two machine translation approaches on a new domain. In *LREC'04, 4th International Conference on Language Resources and Evaluation.*