Augmenting a statistical machine translation baseline system with syntactically motivated translation examples

Elena Irimia, Alexandru Ceaușu
Research Institute for Artificial Intelligence, Romanian Academy
Calea 13 Septembrie no 13, Bucharest, Romania
{elena,aceausu}@racai.ro

Abstract

This paper describes a series of machine translation experiments with the English-Romanian language pair. The experiments were intended to test and prove the hypothesis that syntactically motivated long translation examples added to a base-line 3gram statistically extracted phrase table improves the translation performance in terms of the score BLEU. Extensive tests with a couple of different scenarios were performed: 1) simply concatenating the "extra" translations example to the baseline phrase-table; 2) computing and taking into account perplexities for the POS-string associated to the translation examples; 3) taking into account the number of words in each member of a translation example; 4) filtering the "extra" translation examples by taking into account a score that appreciates the correctness of their lexical alignment. Different combinations of the four scenarios were also tested. Also, the paper presents a method for extracting syntactically motivated translation examples using the dependency linkage of both the source and target sentence. To decompose the source/target sentence into fragments, we identified two types of dependency link-structures - super-links and chains - and used these structures to set the translation example borders.

1. Introduction

Corpus-based paradigm in machine translation has seen various approaches for the task of constructing reliable translation models,

- starting from the naïve "word-to-word" correspondences solution which was studied in the early works (Gale and Church, 1991; Melamed, 1995).
- continuing with the chunk-bounded n-grams (Kupiec, 1993; Kumano and Hirakawa, 1994; Smadja at al., 1996) which were supposed to account for compounding nouns, collocations or idiomatic expressions,
- passing through the early approach of the bounded-length n-grams IBM statistical translation models and the following phrase-based statistical translation models (Och et al, 1999; Marcu and Wong, 2002; etc.),
- exploring the dependency-linked n-grams solutions which can offer the possibility of extracting long and sometimes non-successive examples and are able to catch the structural dependencies in a sentence (e.g., the accord between a verb and a noun phrase in the subject position), see (Yamamoto and Matsumoto, 2003),
- and ending with the double-sided option for the sentence granularity level, which can be appealing since the sentence boundaries are easy to identify but brings the additional problem of fuzzy matching and complicated mechanisms of recombination.

Several studies were dedicated to the impact of using syntactical information in the phrase extraction process over the translation accuracy. Analyzing by comparison the constituency-based model and the dependency based model, Hearne et al. (2008) concluded that "using dependency annotation yields greater translation quality than constituency annotation for PB-SMT". But, as previous works (Groves and Way, 2005; Tinsey et al., 2007) have noted, the new phrase models, created by incorporating linguistic knowledge, do not necessarily improve the translation accuracy by themselves, but in combination with the "old–fashioned" bounded-length phrase models.

The process of extracting syntactically motivated translation examples varies according to the different resources and tools available for specific research groups and specific language pairs. In a detailed report over the syntactically-motivated approaches in SMT, focused on the methods that use the dependency formalism, Ambati (2008) distinguishes the situations when dependency parsers are used for both source and target languages from those in which only a parser for the source side is available. In the latter case, a direct projection technique is usually used to do an annotation transfer from the source to the target translation unit. This approach is motivated by the direct correspondence assumption (DCA, Hwa et al., 2005), that states that dependency relations are preserved through direct projection. The projection is based on correspondences between the words in the parallel sentences, obtained through the lexical alignment (also called word alignment) process. Obviously, the quality of the projection is dependant of the lexical alignment quality. Furthermore, Hwa (2005) notes that the target syntax structure obtained through direct projection is isomorphic to the source syntax structure, thus producing isomorphic translation models. This phenomenon is rarely corresponding to a real isomorphism between the two languages involved.

In the experiments we describe in this paper, we had the advantage of a probabilistic non-supervised dependency analyzer which depends on the text's language only through a small set of rules designed to filter the previously identified links. As both source and target dependency linking analysis is available, there is no need of direct projection in the translation examples extraction and the problem of the "compulsory isomorphism" is avoided.

2. Research background

In previous experiments with an example-based approach on machine translation for the English-Romanian language pair, we developed a strategy for extracting translation examples using the information provided by a dependency-linker described in (Ion, 2007). We then justified our opting for the dependency-linked n-grams approach based on the assumption in (Cranias et al., 1994) that the EBMT potential should rely on exploiting text fragments shorter than the sentence and also on the intuition that a decomposition of the source sentence in “coherent segments”, with complete syntactical structure, would be “the best covering” of that sentence.

The dependency-linker used is based on Yuret’s Lexical Attraction Model (LAM, Yuret, 1998), in whose vision the lexical attraction is a probabilistic measure of the combining affinity between two words in the same sentence. Applied to machine translation, the lexical attraction concept can serve as a mean of guaranteeing the translation examples usefulness. If two words are “lexically attracted” to one another in a sentence, the probability for them to combine in future sentences is significant. Therefore, two or more words from the source sentence that manifest lexical attraction together with their translations in the target language represent a better translation example than a bounded length n-gram.

The choice for the Yuret’s LAM as the base for the dependency analyzer application was motivated by the lack of a dependency grammar for Romanian. The alternative was to perform syntactical analysis based on automatically inducted grammatical models. A basic request for the construction of this type of models is the existence of syntactically annotated corpora from which machine learning techniques could extract statistical information about the ways in which syntactical elements combine. As no syntactically annotated corpus for Romanian was available, the fact that Yuret’s method could use LAM for finding dependency links in a not-annotated corpus made this algorithm a practical choice.

LexPar (Ion, 2007), the dependency links analyzer we used for the experiments described in this paper, is extending Yuret’s algorithm by a set of syntactical rules specific to the processed languages (Romanian and English) that constraints the link formation. It also contains a simple generalization mechanism for the link properties, which eliminates the initial algorithm inadaptability to unknown words. However, the LexPar algorithm does not guarantee a complete analysis, because the syntactic filter can contain rules that forbid the linking of two words in a case in which this link should be allowed. The rules were designed by the algorithm’s author based on his observations of the increased ability of a certain rule to reject wrong links, with the risk of rejecting good links in few cases.

In our research group, significant efforts were invested in experimenting with statistical machine translation methodologies, focused on building accurate language resources (the larger the better) and on fine-tuning the statistical parameters. The aim was to demonstrate that, in this way, acceptable MT prototypes can be quickly developed and the claim was supported by the encouraging Bleu scores we obtained for the Romanian<->English translation system. The translation experiments employed the MOSES toolkit, an open source platform for development of statistical machine translation systems (see next section). The major rationale for selecting this environment was its novel decoding component that facilitates the usage of multiple (factored) translation models.

One of the goals of this paper is to report our findings on the impact of incorporating syntactic information in the translation model by means of a probabilistic dependency link analyzer. Although the non-supervised nature of the analyzer is affecting its recall, using this tool brings the advantage of having syntactic information available for translation without the need of syntactically annotated corpora. We feed the Moses decoder with the new translation model and we compare the translation results with the results of the baseline system.

3. A baseline Romanian-English Machine Translation System

The corpus. The Acquis Communautaire is the total body of European Union (EU) law applicable in the EU Member States. This collection of legislative text changes continuously and currently comprises texts written between the 1950s and 2008 in all the languages of EU Member States. A significant part of these parallel texts have been compiled by the Language Technology group of the European Commission's Joint Research Centre at Ispra into an aligned parallel corpus, called JRC-Acquis (Steinberger et al., 2006), publicly released in May 2006. Recently, the Romanian side of the JRC-Acquis corpus was extended up to a size comparable with the dimensions of other language-parts (19,211 documents)).

For the experiments described in this paper, we retained only 1-1 alignment pairs and restricted the selected pairs so that none of the sentences contained more than 80 words and that the length ratio between sentence-lengths in an aligned pair was less than 7. Finally, the Romanian-English parallel corpus we used contained about 600,000 translation units.

Romanian and English texts were processed based on the RACAI tools (Tufiş et al, 2008) integrated into the linguistic web-service platform available at http://nlp.racai.ro/webservices. After tokenization, tagging and lemmatization, this new information was added to the XML encoding of the parallel corpora. Figure 1 shows the representation of the Romanian segment encoding for the translation unit displayed in Figure 1. The tagsets used were compliant with the MULTEXT-East specifications Version3 (Erjavec, 2004) (for the details of the morpho-syntactic annotation, see http://nl.ijs.si/ME/V3/msd/).

```
<tu id="3936">
    ...
        <seg lang="ro">
            <s id="31985L0337.n.83.1">
                <w lemma="informație"
          ana="Ncfpry">Informațiile</w>
                <w lemma="culege"
          ana="Vmp--pf">culese</w>
                <w lemma="conform"
          ana="Spsd">conform</w>
                <w lemma="art." ana="Yn">art.</w>
                <w lemma="5" ana="Mc">5</w>
                <c>,</c>
                <w lemma="6" ana="Mc">6</w>
                <w lemma="şi" ana="Crssp">şi</w>
                <w lemma="7" ana="Mc">7</w>
                <w lemma="trebui"
        ana="Vmip3s">trebuie</w>
                <w lemma="să" ana="Qs">să</w>
                <w lemma="fi" ana="Vasp3">fie</w>
                <w lemma="lua"
        ana="Vmp--pf">luate</w>
                <w lemma="în" ana="Spsa">în</w>
                <w lemma="considerare"
        ana="Ncfsrn">considerare</w>
                <w lemma="în cadrul" ana="Spcg">în
        cadrul</w>
                <w lemma="procedură"
        ana="Ncfsoy">procedurii</w>
                <w lemma="de" ana="Spsa">de</w>
                <w lemma="autorizare"
        ana="Ncfsrn">autorizare</w>
                <c>.</c>
            </s>
        </seg>
      ...
</tu>
```

Figure 1: Linguistically analysed sentence (Romanian) of a translation unit of the JRC-Acquis parallel corpus

Based on the monolingual data from the JRC-Acquis corpus we built language models for each language. For Romanian we used the TTL (Ion, 2007) and METT (Ceauşu, 2006) tagging modelers. Both systems are able to perform tiered tagging (Tufiş, 1999), a morpho-syntactic disambiguation method that was specially designed to work with large (lexical) tagsets.

In order to build the translation models from the linguistically analyzed parallel corpora we used GIZA++ (Och and Ney, 2000) and constructed unidirectional translation models (EN-RO, RO-EN) which were subsequently combined. After that step, the final translation tables were computed. The processing unit considered in each language was not the word form but the string formed by its lemma and the first two characters of the associated morpho-syntactic tag (e.g. for the wordform "informațiile" we took the item "informație/Nc"). We used for each language 20 iterations (5 for Model 1, 5 for HMM, 1 for THTo3, 4 for Model3, 1 for T2To4 and 4 for Model4). We included neither Model 5 nor Model 6, as we noticed a degradation of the perplexities of the alignment models on the evaluation data.

The MOSES toolkit (Koehn et al., 2007) is a public domain environment, which was developed in the ongoing European project EUROMATRIX, and allows for rapid prototyping of Statistical Machine Translation systems. It assists the developer in constructing the language and translation models for the languages he/she is concerned with and by its advanced factored decoder and control system ensures the solving of the fundamental equation of the Statistical Machine Translation in a noisy-channel model:

$$\text{Target}^* = \text{argmaxTarget } P(\text{Source}|\text{Target})*P(\text{Target}) \quad (1)$$

The P(Target) is the statistical representation of the (target) language model. In our implementation, a language model is a collection of prior and conditional probabilities for unigrams, bigrams and trigrams seen in the training corpus. The conditional probabilities relate lemmas and morpho-syntactic descriptors (MSD), word-forms and lemmas, sequences of two or three MSDs. The P(Source|Target) is the statistical representation of the translation model and it consists of conditional probabilities for various attributes characterizing equivalences for the considered source and target languages (lemmas, MSDs, word forms, phrases, dependencies, etc). The functional argmax is called a decoder and it is a procedure able to find, in the huge search space P(Source|Target)*P(Target) corresponding to possible translations of a given Source text, the Target text that represent the optimal translation, i.e. the one which maximizes the compromise between the faithfulness of translation (P(Source|Target)) and the fluency/grammaticality of the translation (P(Target)). The standard implementation of a decoder is essentially an A* search algorithm.

The current state-of-the-art decoder is the factored decoder implemented in the MOSES toolkit. As the name suggests, this decoder is capable of considering multiple information sources (called factors) in implementing the argmax search. What is extremely useful is that the MOSES environment allows a developer to provide the MOSES decoder with language and translation models externally developed, offering means to ensure the conversion of the necessary data structures into the expected format and further improve them. Once the statistical models are in the prescribed format, the MT system developer may define his/her own factoring strategy. If the information is provided, the MOSES decoder can use various factors (attributes) of each of the lexical items (words or phrases): occurrence form, lemmatized form, associated part-of-speech or morpho-syntactic tag. Moreover, the system allows for integration of higher order information (shallow or even deep parsing information) in order to improve the output lexical items reordering. For further details on the MOSES Toolkit for Statistical Machine Translation and its tuning, the reader is directed to the EUROMATRIX project web-page http://www.euromatrix.net/ and to the download web-page http://www.statmt.org/moses/.

4. Extracting translation examples from corpora (ExTrAct)

In our approach, based on the availability of a dependency-linker for both the source and the target language, the task of extracting translation examples from a corpus contains two sub-problems: dividing the source and target sentences into fragments and setting correspondences between the fragments in the source sentence and their translations in the target sentence. The last problem is basically fragment alignment and we solved it through a heuristic based on lexical alignments produced by GIZA++. The remaining problem was addressed using the information provided by LexPar, the dependency linker mentioned above. With a recall of 60,70% for English, LexPar was considered an appropriate starting point for the experiments (extending or correcting the set of rules incorporated as a filter in LexPar can improve its recall).

Using MtKit, a tool specially designed for the visualization and correction of lexical alignments adapted to allow the graphical representation of the dependency links, we could study the dependency structures created by the identified links inside a sentence and we were able to observe some patterns in the links' behavior: they tend to group by nesting and to decompose the sentence by chaining. Of course, these patterns are direct consequences of the syntactical structures and rules involved in the studied languages, but the visual representation offered by MtKit simplified the task of formalization and heuristic modeling (see Fig. 2).

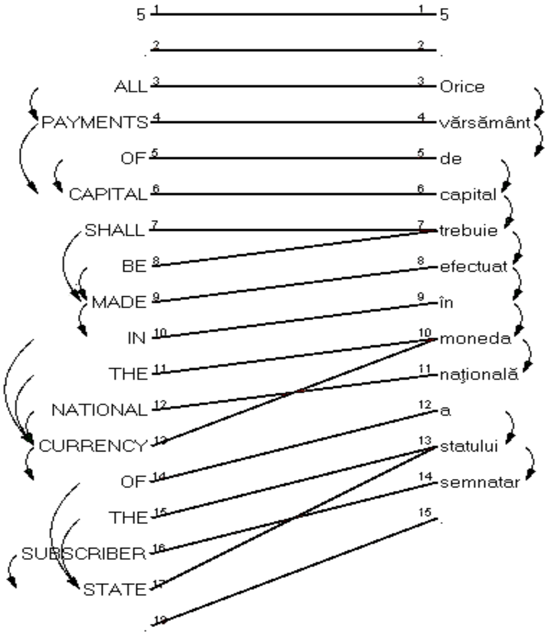
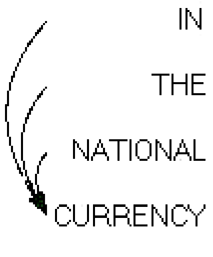


Figure 2. MtKit visualisation of the alignments and links for an English-Romanian translation unit. An arrow marks the existence of a dependency link between the two words it unites. The arrow direction is not relevant for the dependency link orientation.

These properties suggest more possible decompositions for the same sentence, and implicitly the extraction of substrings of different length that satisfy the condition of lexical attraction between the component words.



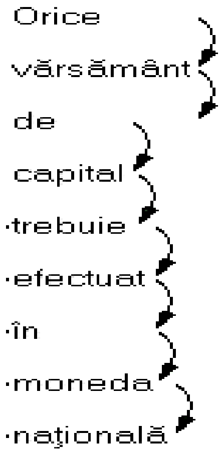


Fig. 3. Superlink structure Fig 4. Chain structure.

Example 1: in Figure 1, from the word sequence "made in the national currency" the flowing subsequences can be extracted: "national currency", "the national currency", „in the national currency", „made in the national currency". The syntactically incomplete sequences and those susceptible of generating errors (like "the national", "in the", "made in the national") are ignored.

The patterns observed above were formalized as superlinks (link structures composed of at least two simple links which nest, see Figure 3) and as chains (link structures composed of at least two simple links or superlinks which form a chain, see Figure 4).

As input data, ExTract (the application that extracts translation examples from corpora) receives the processed corpus and a file containing the lexical alignments produced by GIZA++ (Och and Ney, 2000). We will describe the extracting procedure for a single translation unit U in the corpus, containing Ss (a source sentence) and its translation Ts (a target sentence). In each member (source or target) of the translation unit we identify and extract every possible chaining of links and superlinks, with the condition that the number of chain loops is limited to 3.The limitation was introduced to avoid overloading the database. Subsequent experiments showed that increasing the limitation to 4 or 5 chains did not significantly improve the BLEU score of the translation system. Two list of candidate sentence fragments, from Ss and Ts, are extracted.

Every fragment in both sentences is projected through lexical alignment in a word string (note that this is not the direct syntactical structure projection discussed above, but a surface string projection) in a fragment of the correspondent sentence. *Example: In Figure2, the projection of the target structure "in the national currency" to the source sentence does not involve the dependency link structure's transfer to the source fragment: "în moneda națională". The projection means only a translation correspondence between the source/target word sequences, identified by means of the lexical alignment.*

A projected string of a candidate fragment in Ss is not necessarily part of the list of candidate sentence fragments Ts, and vice versa (sometimes, LexPar is not able to identify all the dependency links in a sentence and the lexical alignments are also subject to errors). But if a fragment candidate from Ss projects to a fragment candidate from Ts (as is the case in our example: "in the national currency" is a superlink candidate from Ss, while "*în moneda națională*" is a chain candidate form Ts), the pair has a better probability of representing a correct translation example. In this stage, the application extracts all the possible translation examples (<source fragment candidate, projected word string>, <projected word string, target fragment candidate>) but distinguish between them, associating a "trust" flag f="2" to the translation examples of the form <source fragment candidate, target fragment candidate>, and a flag f="1" to all the other. Thereby, it is possible to experiment with translation tables of different sizes and different quality levels.

5. Experiments and results

Taking into account results from previous works (Ambati, 2008; Hwa et al., 2005) that proved that dependency-based translation models give improved performance in combination with a phrase-based translation model, we decided to conduct our experiments in a mixed frame: we extracted from the dependency-based translation model only the translation examples longer than (3 source words <-> 3 target words), creating a reduced dependency-based translation model and we combined it with the phrase-based translation model generated with the Moses toolkit.

Starting from the reduced D-based translation model, we can develop two different translation tables, based on the “trust” flags we introduced before:

- a trustful D-based translation table (if we keep only the examples with the flag f=”2”)
- a relaxed D-based translation table (if we accept all the examples, irrespective of the flags).

For the filtering of the D-based translation model we also implemented a heuristic to evaluate the lexical alignment correctness of each translation example. This brought an increase of around 1% (from 52% to 53% for English-Romanian) in the BLEU score.

In an effort to assure the correctness of the examples used by the Moses decoder from the D-based translation model, we introduced a perplexity score which evaluates the MSD-sequence associated to a string against a MSD-language model. Perplexities were computed for both the English and Romanian side of the translation example database. Nor introducing the perplexity scores as translation factors in the decoder, neither filtering the examples in the D-based translation model produced significant difference in the translation performance.

We also wanted to test if we can increase the performance by introducing a score that favors the longer translation examples in the sentence decomposition. Unexpectedly, the results were not improved: the score BLEU was a little bit lower (e.g. a decrease of around 0.3% for English-Romanian, with no statistic relevance). We think this can be explained by the idea that the longer word sequences in the translation are breaking the integrity of the surrounding sequences: the entire sentence translation performance is remaining similar, since the improvements brought by the longer sequences are balanced by the translation errors coming for the shorted sequences. We also assume this effect is noticeable only for the systems in which the base-line translation model already produces good or very-good translations (in our case, a BLEU score of 0.53 for the Moses table is a very good performance).

As we previously mentioned, the initial working corpus contained around 600,000 translation units. From this number, 600 were extracted for tuning and testing. The tuning of the factored translation decoder (the weights on the various factors) was based on the 200 development sentence pairs and it was done using MERT (Och, 2003) method. The testing set contains 400 translation units.

The evaluation tool was the last version of the NIST official mteval script[1] which produces BLEU and NIST scores. For the evaluation, we lowered the case in both reference and automatic translations. The results are synthesized in the following table, where you can notice that our assumption that the trustful table would produce better results than the relaxed one was contradicted by evidence. We thus learned that a wider range of multi-word examples is preferable to a restricted one, even if their correctness was not guaranteed by the syntactical analysis.

		English to Romanian	**Romanian to English**
Moses phrase table	Nist	8.6671	10.7655
	Bleu	*0.5300*	*0.6102*
Dependency trustful table	Nist	8.4998	10.3122
	Bleu	*0.5006*	*0.5812*
Dependency relaxed table	Nist	8.5978	10.3080
	Bleu	*0.5208*	*0.5921*
D-filtered alignment table	Nist	8.6900	10.3235
	Bleu	***0.5334***	***0.6191***
D-filtered align + ppl table	Nist	8.6827	10.1432
	Bleu	*0.5312*	*0.6050*
D-fitered align+ ppl+length table	Nist	8.5000	10.2910
	Bleu	*0.5306*	0.6083

Table 1: Evaluation of the dependency translation table compared with the translation table generated with Moses (on unseen data).

As the scores in the previous table differ only in a superficial manner, we wanted to look closer at the translation results and study how the augmenting of the translation table with new, longer examples, actually affected the translation quality. In a set of 400 sentences, only 93 (~25%) were translated using 1 or more sequences longer than 3 words (i.e. sequences form D-based translation model). When we examined these sentences, we found out that:

- in 15% of them, using the D-based sequences had a negative impact on the BLEU score (but not necessary on the quality of the translation as assessed by a human evaluator);
- in 50% of the cases, the final form of the translation didn’t change (no effect on the performance) ;
- in 35% of the cases, the quality of the translation improved in terms of both the BLEU score and the human evaluator opinion.

Example 2: In the following example the reader can notice a case in which the n-gram matching (and consequently the BLEU score) between the translation and the Romanian reference are improved in the D-based model BEST TRANSLATION (see the bolded words). The example contains also a case in which the performance is not affected by the use of a longer translation example (see the second italic text fragment in the TRANSLATION HYPOTHESIS DETAILS[2]).

English reference (source):
member states shall adopt the measures necessary to comply with this directive within six months of its notification and shall forthwith inform the commission thereof .

[1] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v12.pl

[2] Translation hypothesis details as outputted by the Moses decoder.

Romanian reference (target):
*statele membre **iau măsurile necesare** pentru a se conforma prezentei directive în termen de şase luni de_la data notificării acesteia şi informează imediat comisia cu_privire_la aceasta .*

Moses BEST TRANSLATION: *statele membre adoptă măsurile necesare pentru a se conforma prezentei directive în termen de şase luni de_la notificarea acesteia şi informează de_îndată comisia cu_privire_la aceasta.*
TRANSLATION HYPOTHESIS DETAILS:
SOURCE: [0..2] member states shall
TRANSLATED AS: statele membre
SOURCE: [3..3] adopt
*TRANSLATED AS: **adoptă***
SOURCE: [4..5] the measures
TRANSLATED AS: măsurile
SOURCE: [6..7] necessary to
TRANSLATED AS: necesare pentru a
SOURCE: [8..10] comply with this
TRANSLATED AS: se conforma prezentei
SOURCE: [11..12] directive within
TRANSLATED AS: directive în termen
SOURCE: [13..14] six months
TRANSLATED AS: de şase luni
SOURCE: [15..17] of its notification
TRANSLATED AS: de_la notificarea acesteia
SOURCE: [18..18] and
TRANSLATED AS: şi
SOURCE: [19..21] shall forthwith inform
TRANSLATED AS: informează de_îndată
SOURCE: [22..24] the commission thereof
TRANSLATED AS: comisia cu_privire_la aceasta
SOURCE: [25..25] .
TRANSLATED AS: .

SOURCE/TARGET SPANS:
SOURCE: 0-1-2 3 4-5 6-7 8-9-10 11-12 13-14 15-16-17 18 19-20-21 22-23-24 25
TARGET: 0-1 2 3 4-5-6 7-8-9 10-11-12 13-14-15 16-17-18 19 20-21 22-23-24 25

D-based model BEST TRANSLATION: *statele membre iau măsurile necesare pentru a se conforma prezentei directive în termen de şase luni de_la notificarea acesteia şi informează de_îndată comisia cu_privire_la aceasta.*
TRANSLATION HYPOTHESIS DETAILS:
SOURCE: [0..5] member states shall adopt the measures
*TRANSLATED AS: statele member **iau** măsurile*
SOURCE: [6..7] necessary to
TRANSLATED AS: necesare pentru a
SOURCE: [8..10] comply with this
TRANSLATED AS: se conforma prezentei
SOURCE: [11..12] directive within
TRANSLATED AS: directive în termen
SOURCE: [13..14] six months
TRANSLATED AS: de şase luni
SOURCE: [15..17] of its notification
TRANSLATED AS: de_la notificarea acesteia
SOURCE: [18..18] and
TRANSLATED AS: şi
SOURCE: [19..23] shall forthwith inform the commission
TRANSLATED AS: informează de_îndată comisia
SOURCE: [24..25] thereof .
TRANSLATED AS: cu_privire_la aceasta .

SOURCE/TARGET SPANS:
SOURCE: 0-1-2-3-4-5 6-7 8-9-10 11-12 13-14 15-16-17 18 19-20-21-22-23 24-25
TARGET: 0-1 2-3 4-5-6 7-8-9 10-11-12 13-14-15 16-17-18 19 20-21-22 23-24-25

Example 3: This example presents a case in which the score Bleu of the D-based translation is decreased (a case in the 15% of negative impact mentioned before), but the translation remains very good for the human evaluator's perspective (see the bolded fragments).

English reference (source):
for the purpose of determining entitlement to benefits in kind pursuant to article 22 (1) (a) and article 31 of the regulation , " member of the family " means any person regarded as a member of the family under the law on the public health service .
Romanian reference (target):
*în scopul determinării dreptului la prestații în natură în aplicarea art. 22 alin. (1) lit. (a) şi a art. 31 din regulament , " membru de familie " reprezintă orice persoană considerată membru de familie **în_conformitate_cu legea privind serviciul public de sănătate** .*

*Moses BEST TRANSLATION: în scopul determinării dreptului la prestații în natură în temeiul articolului 22 alineatul (1) litera (a) şi articolul 31 din regulament , " membru de familie " reprezintă orice persoană considerată membru de familie **în_conformitate_cu legea privind sănătatea_publică** .*

TRANSLATION HYPOTHESIS DETAILS:
SOURCE: [0..2] for the purpose
TRANSLATED AS: în scopul
SOURCE: [3..5] of determining entitlement
TRANSLATED AS: determinării dreptului
SOURCE: [6..7] to benefits
TRANSLATED AS: la prestații
SOURCE: [8..9] in kind
TRANSLATED AS: în natură
SOURCE: [10..12] pursuant to article
TRANSLATED AS: în temeiul articolului
SOURCE: [13..13] 22
TRANSLATED AS: 22
SOURCE: [14..15] (1
TRANSLATED AS: alineatul (1
SOURCE: [16..17]) (
TRANSLATED AS:) litera (
SOURCE: [18..19] a)
TRANSLATED AS: a)
SOURCE: [20..22] and article 31
TRANSLATED AS: şi articolul 31
SOURCE: [23..25] of the regulation
TRANSLATED AS: din regulament
SOURCE: [26..28] , " member
TRANSLATED AS: , " membru
SOURCE: [29..31] of the family

TRANSLATED AS: de familie
SOURCE: [32..32] "
TRANSLATED AS: "
SOURCE: [33..35] means any person
TRANSLATED AS: reprezintă orice persoană
SOURCE: [36..38] regarded as a
TRANSLATED AS: considerată
SOURCE: [39..39] member
TRANSLATED AS: membru
SOURCE: [40..42] of the family
TRANSLATED AS: de familie
SOURCE: [43..43] under
TRANSLATED AS: în_conformitate_cu
SOURCE: [44..45] the law
TRANSLATED AS: legea
SOURCE: [46..47] on the
TRANSLATED AS: privind
SOURCE: [48..49] public health
TRANSLATED AS: sănătatea_publică
SOURCE: [50..51] service .
TRANSLATED AS: .

SOURCE/TARGET SPANS:
SOURCE: 0-1-2 3-4-5 6-7 8-9 10-11-12 13 14-15 16-17 18-19 20-21-22 23-24-25 26-27-28 29-30-31 32 33-34-35 36-37-38 39 40-41-42 43 44-45 46-47 48-49 50-51
TARGET: 0-1 2-3 4-5 6-7 8-9-10 11 12-13-14 15-16-17 18-19 20-21-22 23-24 25-26-27 28-29 30 31-32-33 34 35 36-37 38 39 40 41 42

D-based model BEST TRANSLATION: în scopul determinării dreptului la prestații în natură în temeiul articolului 22 alineatul (1) litera (a) şi articolul 31 din regulament , " membru de familie " reprezintă orice persoană considerată membru de familie , ***conform legislației în domeniul sănătății_publice*** *.*

TRANSLATION HYPOTHESIS DETAILS:
SOURCE: [0..5] for the purpose of determining entitlement
TRANSLATED AS: în scopul determinării dreptului
SOURCE: [6..9] to benefits in kind
TRANSLATED AS: la prestații în natură
SOURCE: [10..12] pursuant to article
TRANSLATED AS: în temeiul articolului
SOURCE: [13..13] 22
TRANSLATED AS: 22
SOURCE: [14..15] (1
TRANSLATED AS: alineatul (1
SOURCE: [16..17]) (
TRANSLATED AS:) litera (
SOURCE: [18..20] a) and
TRANSLATED AS: a) şi
SOURCE: [21..21] article
TRANSLATED AS: articolul
SOURCE: [22..25] 31 of the regulation
TRANSLATED AS: 31 din regulament
SOURCE: [26..28] , " member
TRANSLATED AS: , " membru
SOURCE: [29..31] of the family
TRANSLATED AS: de familie
SOURCE: [32..32] "
TRANSLATED AS: "
SOURCE: [33..35] means any person
TRANSLATED AS: reprezintă orice persoană
SOURCE: [36..38] regarded as a
TRANSLATED AS: considerată
SOURCE: [39..39] member
TRANSLATED AS: membru
SOURCE: [40..42] of the family
TRANSLATED AS: de familie
SOURCE: [43..45] under the law
TRANSLATED AS: , conform legislației
SOURCE: [46..47] on the
TRANSLATED AS: în
SOURCE: [48..49] public health
TRANSLATED AS: domeniul sănătății_publice
SOURCE: [50..51] service .
TRANSLATED AS: .

SOURCE/TARGET SPANS:
SOURCE: 0-1-2-3-4-5 6-7-8-9 10-11-12 13 14-15 16-17 18-19-20 21 22-23-24-25 26-27-28 29-30-31 32 33-34-35 36-37-38 39 40-41-42 43-44-45 46-47 48-49 50-51
TARGET: 0-1-2-3 4-5-6-7 8-9-10 11 12-13-14 15-16-17 18-19-20 21 22-23-24 25-26-27 28-29 30 31-32-33 34 35 36-37 38-39-40 41 42-43 44

6. Conclusion

We briefly presented only a small part of the various machine translation experiments done in the last year in our research group (including both statistical and dependency-based translation models, the language pair English-Romanian and other languages like Greek and Slovene). We tried to look for solutions to improve the already very good performance of the baseline system on the Romanian-English pair, but in terms of the automatic evaluation method we used (the BLEU/NIST score), the results were not convincing. We analyzed and discovered that the performance increasing impact of adding longer dependency-motivated translation examples can be observed in 5% percent of the translated sentences. We assume that the expected important increasing in the system's performance was not to be seen because the translation quality offered by the baseline MOSES configuration was already very good. Future experiments should address other domains and literary registries, with lesser baseline performances, to check our assumption.

7. Acknowledgements

The work reported here is funded by the STAR project, financed by the Ministry of Education, Research and Innovation under the grant no 742 and by the ACCURAT project funded from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 248347.

8. References

Ambati, V. 2008. Dependency Structure Trees in Syntax Based Machine Translation, 11-734 Spring 2008, Survey Report, http://www.cs.cmu.edu/~vamshi/publications/DependencyMT_report.pdf

Ceausu Al. 2006. Maximum Entropy Tiered Tagging. In Janneke Huitink & Sophia Katrenko (editors), Proceedings of the Eleventh ESSLLI Student Session, pp. 173-179

Cranias, L., H. Papageorgiou and S. Piperidis. 1994. A Matching Technique in Example-Based Machine Translation. In Proceedings of the 15th conference on Computational linguistics - Volume 1, Kyoto, Japan 100–104.

Erjavec, T. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04, ELRA, Paris, pp. 1535 – 1538

Gale, W. and K. Church, 1991. Identifying Word Correspondences in Parallel Texts. In Proceedings of the 4th DARPA Speech and Natural Language Workshop, Pacific Grove, CA., pp. 152-157.

Groves D. & Way A. 2005. Hybrid Example-Based SMT: the Best of Both Worlds? In Proceedings of ACL 2005 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond, p. 183–190, Ann Arbor, MI.

Hearne, M., S. Ozdowska, and J. Tinsley. 2008. Comparing Constituency and Dependency Representations for SMT Phrase-Extraction. In Proceedings of TALN '08, Avignon, France.

Hwa, R., Ph. Resnik, A. Weinberg, C. Cabezas and O. Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. Nat. Lang. Eng., 11(3):311–325, September.

Ion, R. 2007. Word Sense Disambiguation Methods Applied to English and Romanian, PhD thesis (in Romanian), Romanian Academy, Bucharest, 138 p.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Wade, S., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. 2007. MOSES: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.

Kumano, A. and H. Hirakawa. 1994. Building an MT dictionary from parallel texts based on linguistic and statistical information. In COLING-94: Proceedings of the 15th International Conference on Computational Linguistics, Kyoto, Japan, pages 76-81.

Kupiec, J. 1993. An Algorithm for Finding Noun Phrases Correspondences in Bilingual Corpora. In 31st Annual Meeting of the Association for Computational Linguistics, Columbus, OH., pages 23-30.

Marcu D. and W. Wong. 2002. A Phrased-Based, Joint Probability Model for Statistical Machine Translation. In Proceedings Of the Conference on Empirical Methods in Natural Language Processing (EMNLP 02); pages 133-139, Philadelphia, PA, July.

Melamed, I.D. 1995. Automatic Evaluation and Uniform Filter Cascades for Inducing N-best translation lexicons. In proceedings of the Third Annual Workshop on Very Large Corpora, Cambridge, England, pp. 184-198.

Och, F. J. 2003. Minimal Error Rate Training in Statistical Machine Translation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, July 2003, pp. 160-167.

Och, F.J., Ney H. 2000. Improved Statistical Alignment Models. In Proceedings of the 38th Conference of ACL, Hong Kong, pp. 440-447

Och, F.-J., Ch. Tillmann and H. Ney. 1999. Improved Alignment Models for Statistical Machine Translation. In Proc. of the Joint Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC 99), pages 20–28, College Park, MD, June.

Smadja, F., K.R. McKeown and V. Hatzivassiloglou. 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. Computational Linguistics 22(1):1-38.

Steinberger R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the 5th LREC Conference, Genoa, Italy, 22-28 May, pp. 2142-2147

Tinsley J., Hearne M. & Way A. 2007. Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation. In Proceedings of The Sixth International Workshop on Treebanks and Linguistic Theories (TLT-07), Bergen, Norway.

Tufiş, D. 1999. Tiered Tagging and Combined Language Models Classifiers. In Václav Matousek, Pavel Mautner, Jana Ocelíková, and Petr Sojka, editors, Text, Speech and Dialogue (TSD 1999), Lecture Notes in Artificial Intelligence 1692, Springer Berlin / Heidelberg,. ISBN 978-3-540-66494-9, pp. 28-33.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2008). RACAI's Linguistic Web Services. In Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008, Marrakech, Morocco. ELRA - European Language Ressources Association. ISBN 2-9517408-4-0.

Yamamoto, K. and Y. Matsumoto. 2003. Extracting translation knowledge from parallel corpora. In: Michael Carl & Andy Way (eds.) Recent advances in example-based machine translation (Dordrecht: Kluwer Academic Publishers, 2003); pages 365-395.

Yuret, D. 1998. Discovery of linguistic relations using lexical atrraction. PhD thesis, Department of Computer Science and Electrical Engineering, MIT.