PARALLEL CORPORA, ALIGNMENT TECHNOLOGIES AND FURTHER PROSPECTS IN MULTILINGUAL RESOURCES AND TECHNOLOGY INFRASTRUCTURE

Dan TUFIŞ, Radu ION

Research Institute for Artificial Intelligence, Romanian Academy, 13, "13 Septembrie", 050711, Bucharest, Romania
{tufis, radu}@racai.ro

Multilingual technologies, which to a large extent are language independent, provide a powerful support for easier building of annotated linguistic resources for languages where such resources are scarce or missing. All these technologies require parallel corpora in order to achieve their ends. Parallel texts encode extremely valuable linguistic knowledge because the linguistic decisions made by the human translators in order to faithfully convey the meaning of the source text can be traced and used as evidence on linguistic facts which, in a monolingual context, might be unavailable to or overlooked by a computer program. In this paper we will briefly present some underlying multilingual technologies and methodologies we developed for exploiting parallel corpora and we will discuss their relevance for cross-linguistic annotation transfer and applications.

Key words: Word alignment, annotation, collocation, POS disambiguation, WSD, encoding, parallel corpora, multilingual technologies, tagging, wordnet.

1. INTRODUCTION

In not a too distant past, parallel data was hard to come by. Nowadays, the Internet plays an important role in acquisition of all sorts of data. Among the diversity of written material available, the interested researcher can find lots of parallel texts that more often than not (for some notable exceptions see for instance the JRC Acquis corpus[1] or the Hansards corpus[2]) take the form of multilingual WEB sites. Parallel corpora encode extremely valuable linguistic knowledge in that the linguistic decisions made by the human translators in order to faithfully convey the meaning of the source text, can be traced and used as evidence on linguistic facts which, in a monolingual context, might be unavailable to, or overlooked by a computer program.

The automatic identification in a parallel corpus of the segments of texts that represent reciprocal translations is a prerequisite for taking advantage of the implicit linguistic knowledge embedded into the translations. This problem, known as parallel corpus alignment, can be defined at various levels of text segmentation granularity (paragraph, sentence, phrase, word) with different degrees of difficulty. The computational approaches to the alignment problem range from symbolic/rule-based to purely statistical ones and one could easily find arguments for/against each of these approaches. Lately, with more parallel data available, there is a visible tendency towards statistical approaches, but in general, the best compromise between the expected accuracy and the necessary development efforts can be reached by taking a mix approach. Irrespective of the text segmentation granularity and of the prevalence of either a rationalistic or an empirical model in a mix approach, the larger the parallel corpus, the better the alignment accuracy (obviously, for purely statistical approaches, mass data is even more critical). On the other hand, depending on the alignment granularity, required accuracy, and the purpose of the alignment, the input textual data might need pre-processing steps in all languages of the parallel corpus (e.g. sentence splitting, tokenization,

[1] http://langtech.jrc.it/JRC-Acquis.html
[2] http://www.isi.edu/natural-language/download/hansard/

POS-tagging and lemmatization) or at least in one of the languages of the corpus (e.g. chunking, dependency linking/parsing, and word sense disambiguation).

Two segments of texts from a bitext which represent reciprocal translations make a translation unit. A translation unit may contain, in one or both paired languages, one or more textual units (paragraph, sentence, phrase, word) and one distinguishes between the 1:1 and non-1:1 alignment translation units. While at the paragraph granularity level the non-1:1 alignments are exceptional (most sentence aligners assume the number of paragraphs to be the same in the two sides of a bitext, and thus only 1:1 alignments are considered), at the sentence or phrase level they are quite rare (usually no more than 5-10% of the total number of translation units). At the word level, the non-1:1 alignments are more frequent and their number strongly depends on the language pair and on the type of translation (literal versus free translation). Another source of increased difficulties for fine-grained alignments is that while at the paragraph and (to a large extent) sentence level the ordering of the textual units is preserved in both sides of a bitext (discourse coherence requirement), at the finer grained level this is not true in general (the word or phase ordering being ruled in each language by its syntax).

In this paper we will discuss the use of a mix approach to the alignment problem, focussed on the finest-grained level, the word alignment. Word-alignment technology stemmed from the research in Statistical Machine Translation (SMT), which is the first and best known use of it. Here, we will describe some less advertised but very useful applications of the word-alignment technology in lexical semantics knowledge acquisition and validation and in cross-lingual transfer of linguistic (syntactic and semantic) annotations.

The paper is organised as follows: in section 2 we briefly describe the preliminary pre-processing steps required for a high accuracy word alignment as well as the lexical alignment itself. The 3rd section is dedicated to the use of word aligned parallel texts. First we discuss the problem of the word-sense disambiguation (WSD) and we will show how a WSD application can serve to the validation and updating of aligned wordnets of the EuroWordNet [35] or BalkaNet type [32]. Then, we discuss how word alignment can be used in annotation transfer and review some ongoing syntactic and semantic data transfer projects. Section 4 is devoted to a brief discussion of an ongoing cross-lingual (four languages) investigation of collocation phenomena and section 5 introduces a web-based platform for multilingual NLP services.

2. PREPROCESSING STEPS

If one has at his/her disposal a parallel text with is already aligned at the sentence level, the word alignment procedure has its raw input assured. If this is not the case, the parallel text that is supposed to be aligned at the paragraph level (1:1 alignments between paragraphs), needs to be realigned at sentence level because the granularity of the alignment has to be increased so as to simplify the job of the lexical aligner.

Bob Moore presents a state of the art sentence aligner [18] which uses a three stage hybrid approach. In the first stage, the algorithm uses length-based methods [10] assuming the monotonic ordering of the paragraphs and sentences in the two languages of a bitext. In the second stage, a translation equivalence table is estimated from the aligned corpus resulted in the first stage. The method used for translation equivalents estimation is based on IBM-1 model [4]. The final step uses a combination of length-based methods and word correspondence to find only 1:1 sentence alignments. The aligner has an excellent precision but, apparently, it cannot process more than 100,000 sentence pairs.

We developed a sentence aligner [6] inspired by Moore's program which removes the 1:1 alignment restriction, the assumption on the monotonic ordering of the sentences in the two languages, as well as the upper limit on the number of sentence-pairs that can be aligned. It has a comparable precision but a better recall than Moore's aligner.

The sentence aligner consists of a hypothesis generator which creates a list of plausible sentence alignments from the parallel corpus (see [6] for more details) and a filter which removes the improbable alignments. The filter is an SVM binary classifier [8] initially trained on a Gold Standard. The features of the initial SVM model are: the word sentence length, the number of non-word tokens, and the rank correlation for the first 25% of the most frequent words in the two parts of the training bitext. This model is used to preliminary filter alignment hypotheses generated from the parallel corpus. The set of the pairs of sentences that remained after this filtering is used as the input for an EM algorithm which builds a word translation

equivalence table by a similar approach to the IBM model-1 procedure. The SVM model is rebuilt (again from the Gold Standard) this time including, as an additional feature, the number of word translation equivalents existing in the sentences of a candidate alignment pair. This new model is used by the SVM classifier for the final sentence alignment of the parallel corpus.

Various methods of word alignment have been proposed over the time of which a very simple one, the IBM model 1 ([4]), only requires the parallel text to be aligned at the sentence level granularity in order to produce lexical alignments. Other lexical alignment methods are more demanding in that they need different kinds of annotations of the parallel text such as POS-tagging, lemmatization, shallow parsing, and so on. Our word alignment algorithms, MEBA and YAWA ([26], [28]), require the following preprocessing steps so as to produce lexical alignments:

- ***Text segmentation.*** The first pre-processing step in most NLP systems deals with text segmentation. In our processing chain this step is achieved by a modified version (much faster) of the multilingual tokeniser MtSeg which has tokenization resources for many western European languages, developed within the MULTEXT project, further enhanced in the follow up MULTEXT-EAST project [31] with corresponding resources for Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene. Our segmenter is a part of the Perl module called TTL ([12]) and is able to recognize paragraphs, sentence and clause boundaries, dates, numbers and various fix phrases (which form a core of a future Named Entity recogniser based on regular expressions), and to split clitics or contractions (where the case). We significantly updated the tokenization resources for Romanian and English (the languages we have been most interested in lately).
- ***POS-tagging.*** It is generally known that the accuracy of POS-tagging depends on the quality of the language model underlying the morpho-lexical processing, which, on its turn, is highly dependent on the quality and quantity of the training data and on the tagset of the language model. For languages with a productive inflectional morphology the morpho-lexical feature-value combinations may be very numerous, leading to very large tagsets with unavoidable training data sparseness threats. The lack of sufficient training data affects the robustness of the language models which, consequently, will generate an increased number of tagging errors at the run time. To cope with the tagset cardinality problem we developed the tiered-tagging methodology [30] and implemented it using the TnT trigram HMM tagger [3]. The methodology involves the use of a reduced hidden corpus tagset, automatically constructed from the large targeted lexical tagset, and a procedure to map back the reduced tagset into the large one, used in the final annotated text. The two tagsets (the lexical and corpus tagsets) are related by a subsumption relation. When the reduction of the cardinality of the large tagset is information lossless (redundancy elimination) the mapping from the reduced tagset to the large one is deterministic and it is simply ensured by a lookup of a dictionary. For tagset reduction with information loss, which ensures a much significant reduction of the lexical tagsets, the recovering of the left out morpho-lexical information, although to a large extent deterministic, requires an additional preprocessing to solve some non-deterministic cases. In the previous version of the tiered tagging approach we used several hand-crafted rules (regular expressions defined over the reduced tagset, with a span of ±4 tags around the ambiguously mapped tags). Recently, we have re-implemented the tiered tagging methodology, by relying on a combination between an HMM tagger, called TTL [12], [34], which produces also the lemmatization, and a maximum-entropy tagger [5]. The HMM tagger works with the reduced tagset while the ME-tagger ensures the mapping of the first tagset onto the large one (the lexical tagset) dispensing on the hand-written mapping rules.
- ***Lemmatization*** is in our case a straightforward process, since the monolingual lexicons developed within MULTEXT-EAST contain, for each word, its lemma and the morpho-lexical tag. Knowing the word-form and its associated tag, the lemma extraction is simply a matter of lexicon lookup for those words that are in the lexicon. For the unknown words, which are not tagged as proper names, a set of lemma candidates is generated by a set of suffix-stripping rules induced from the word-form lexicon. A four-gram letter Markov model (trained on lemmas in the word-form dictionary) is used to choose the most likely lemma (see [12] for details).
- ***Chunking***. By means of a set of language dependent regular expressions defined over the tagsets, our chunker accurately recognizes the (non-recursive) noun phrases, adjectival/adverbial phrases,

prepositional phrases and verb complexes (analytical realization of tense, aspect mood and diathesis and phrasal verbs) both for Romanian and English ([12] provides further details).

2.1. Word Alignment

A representation of pairs of words $<w^i_{L1}\ w^j_{L2}>$ that represent mutual translations (called translation equivalence pairs) occurring in the same translation unit is the task of a word alignment algorithm. If w^i_{L1} or w^j_{L2} is NULL, we have a case of *null alignment* where one word in one part of the bitext is not translated in the other part. When w^i_{L1} , w^j_{L2} or both appear in several translation equivalence pairs in the same translation unit, they correspond to *multi-word expression alignments*.

The bitext, pre-processed as described in the previous section, is fed into the word alignment engine, called COWAL [26], [28] which is a wrapper of two stand-alone aligners (YAWA and MEBA). COWAL merges the alignments produced by each stand-alone aligner and then uses a trained SVM classifier to prune the unlikely alignment links. The classifier is based on the LIBSVM kit [8] used with the default parameters (C-SVC classification and radial basis kernel function). The classifier was trained with positive and negative hand-validated examples of word alignment links.

The usefulness of the aligner combination has been convincingly demonstrated on the occasion of the Shared Task on Word Alignment organized by the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond" [17]. We participated (on the Romanian-English track) with the two, standalone aligners and the combined one [28]. Out of the 37 competing systems, COWAL was rated the first, MEBA the 20th and TREQ-AL, (the former version of YAWA), was rated the 21st. Meanwhile, both stand-alone aligners have been improved (see table 1) in various ways and trained on more data, but the combined aligner still performs better than both of them.

Table 1. Combined alignment

Aligner	Precision	Recall	F-measure
YAWA	88.80%	74.83%	81.22%
MEBA	92.15%	73.40%	81.71%
COWAL	87.26%	80.94%	83.98%

COWAL is now embedded into a larger, DOTNET-based platform, called MTkit (see figure 1) that incorporates the tools for bitexts pre-processing, a graphical interface that allows for comparing and editing different alignments (insertion of 1:1 links and multi-word expression alignment links and deletion of links with the possibility of saving this annotation), as well as a word sense disambiguation module (described in the next section).

3. EXPLOITING THE ALIGNMENTS

In what follows, we will show a few examples of the use we have given to the word alignments. The aligned corpora we worked with were:

- the Ro-En sub-corpus of the "1984" multilingual [7] corpus (about 110,000 tokens per language),
- a partial translation in Romanian of SemCor2.0 (about 175,000 tokens per language) and
- the four-language (English-Romanian-French-German) sub-corpus of the 21-language parallel corpus Acquis Communautaire[3] (about 8 million tokens per language).

We were interested only in open-class words the alignment of which is significantly more accurate than the "all words" alignment. This means that most of the alignment errors in the evaluation from table 1, were related to functional words and punctuation.

3.1. Aligned Wordnets Validation

Once we have acquired the translation equivalents, it is reasonable to expect that the words of a translation pair $<w^i_{L1}, w^j_{L2}>$ share at least one conceptual meaning stored in an interlingual sense inventory. In the BalkaNet project [32] we used the Princeton WordNet (PWN) [9] as an interlingual index. Based on

[3] http://langtech.jrc.it/JRC-Acquis.html

the interlingually aligned wordnets, obtaining the sense labels for the words in a translation pair is straightforward [15]:

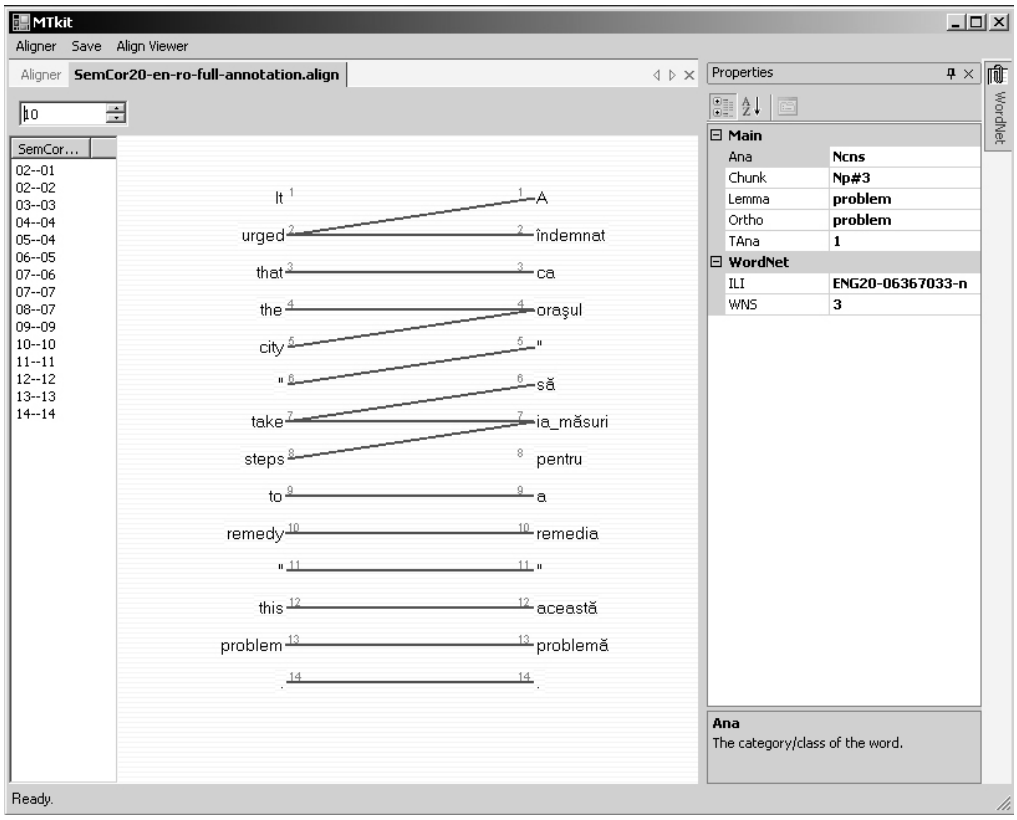

Figure 1. The graphical user interface of the COWAL combined word aligner.

a) identify for w^i_{L1} the synset S^i_{L1} and for w^j_{L2} the synset S^j_{L2} so that S^i_{L1} and S^j_{L2} are projected over the same interlingual concept. The index of this common interlingual concept (*ILI*) is the sense label of the two words w^i_{L1} and w^j_{L2}.

b) if no common interlingual projection is found for the synsets to which w^i_{L1} and w^j_{L2} belong, the senses of the two words will be given by the indexes of the most similar interlingual concepts corresponding to the synsets of the two words. The semantic-similarity score is computed (for nouns and verbs only) as $SYM(ILI_1, ILI_2) = 1/(1+k)$ where k is the number of PWN links from ILI_1 to ILI_2 or from both ILI_1 and ILI_2 to the nearest common ancestor.

In case none of the two cases above holds, then it is very likely that there are some problems which can be categorized as follows:

i) the translation pair is wrong (either because of human translator or because of the word aligner), so it is natural not to find any *ILI* matching for the two words of the pair;

ii) one or both words do not have implemented the relevant senses;

iii) one or both words are missing from the relevant existing synsets;

iv) one or both synsets to which the words of the current translation pair belong are not correctly linked to the relevant *ILI*;

v) the two words in the current translation pair have different POS. Since all the BalkaNet wordnets were aligned to PWN version 2.0 preserving the POS of the synsets, all the cross-POS translations will fit this case.

For the semantic validation of the wordnets created during the BalkaNet project the cases i) and v) were not relevant. The alignment pairs of the ii), iii) and iv) types extracted from the various bilingual sub-corpora of the "1984" corpus were validated by native speakers, with very good command of English. As a result, many synsets were extended with missing literals, the missing synsets were added, and wrong

interlingual projections were corrected. The final report of the BalkaNet[4] project gives a detailed quantitative and qualitative account of the errors and incompleteness that were detected by this procedure (and corrected by each partner).

Since the BalkaNet project finished, we have consistently extended the Romanian wordnet [34] (currently it contains more than 34,500 synsets, and this number is steadily growing). We repeated the semantic validation procedure several times until we haven't noticed any problematic case in the Ro-En sub-corpus of "1984".

3.2. A Romanian Version of SEMCOR

We have applied a slightly modified variant of the wordnets validation procedure on a partial translation of SemCor2.0 into Romanian. The two sets of documents were sentence aligned by the translators (8276 aligned sentence pairs) and word aligned by the COWAL word aligner. Based on a random set of links, we have estimated the alignment correctness for the closed-class categories at 98.5%. With this accuracy, the word alignment error problem didn't have a relevant contribution to a number of problems that showed up.

With the word senses available in the English part, the transfer procedure of these senses to the Romanian part of the corpus proceeded in the following manner (we considered only the translation pairs in which the POS was preserved because of the ILI projections that preserve POS): for each pair of aligned words, we checked whether the ILI number of the sense marked on the English word corresponded to a Romanian synset containing the Romanian translation equivalent or to a hypo/hypernym of it. Because one already knows the meaning of one member of the translation equivalence pair, the procedure of validating aligned wordnets (which is also the procedure we used in implementing a multilingual word sense disambiguation procedure called WSDTool, presented in the next section and in detail in [27]) is thus turned into a semantic labeling transfer procedure with virtually no changes.

Out of the 178499 tokens in the English part[5], only 79595 tokens were sense marked-up but 3535 were not translated into Romanian and 3694 translation pairs had different part of speech. For the remaining 72366 POS-preserving translated English content words, the sense correspondence was found for 48392 Romanian tokens (66.87%). We thoroughly analyzed the 23974 cases of the unmatched senses and found two main situations:

a) Romanian translation equivalent was not in the Romanian wordnet: 11930 cases;
b) The Romanian synset with the ILI of the English word exists, but does not include the Romanian word (incomplete synset): 12044 cases.

While the a) situation was somehow expected (given that the Romanian wordnet contains much less synsets and literals than PWN, leaving room for long-time development efforts for our team), the second situation was worrying and we concentrated our evaluation on that case. As expected, a large number (more than 4000) of the translation equivalence pairs coming under this rubric were alignment errors but still, numerous incomplete synsets (more than 7000) were detected. While a part of this work has been already done, the extension of the remaining synsets with the missing literals is one of our future activities.

3.3. Multilingual WordNet-Based Sense Disambiguation

The WSD task can be stated as being able to associate to an word (w), ambiguous in a text or discourse, the sense (s_k) which is distinguishable from other senses ($s_1 \ldots s_{k-1}\ s_{k+1} \ldots s_n$) and is prescribed for that word by a reference semantic lexicon. The task of word sense disambiguation (WSD) requires one reference sense inventory, in terms of which the senses of the target words will be labeled. We argued at length elsewhere [29] that a meaningful discussion of the performances of a WSD system cannot dispense of clearly specifying the sense inventory it uses, and the comparison between two WSD systems that use different sense inventories is frequently more confusing than illuminating. Essentially, this is because the differences in the semantic distinctions (sense granularities), as used by different semantic dictionaries (sense

[4] http://www.ceid.upatras.gr/Balkanet/resources.htm
[5] The Romanian part contains 175603 tokens.

repositories), make the difficulty of the WSD task range over a large spectrum. For instance, the discrimination of homographs (more often than not having different parts of speech), is much simpler than the metonymic distinctions.

It is straightforward to turn the previous validation procedure into a WSD engine and we claim that state-of-the-art (and even better) performances can be achieved in sense disambiguation of the words in a parallel text, provided the respective bitext is word aligned and aligned wordnets exists for the considered languages.

As mentioned, we used the PWN version 2.0 (PWN2.0) synset identifiers as the reference sense inventory. Through the 1-1 mapping, existing between the synsets in the Romanian wordnet and PWN2.0, the SUMO/MILO [19] and DOMAINS [16] labels became available in our wordnet (as in all the other wordnets which are aligned to PWN2.0). Since both SUMO/MILO and DOMAINS synset labeling is POS insensitive, their use is extremely helpful in assigning senses to the words in a cross-pos translation equivalence pair. For instance, if we consider the En-Ro translation equivalence pairs <*fire*-verb *concediat*[6]-adj> the POS preserving ILI-based alignment between PWN2.0 and RoWN is not really helpful. However, if one uses instead of synset identifiers the SUMO labels, a match would be found since the senses *fire*:4 for the verb and *concediat:1* for the adjective belong to synsets labeled by the same SUMO concept *TerminatingEmployment.* An extra-bonus is that one can infer that the first sense of the adjective *fired* (the translation of *concediat*) is derived from the fourth sense (and no other one) of the verb *fire*. This is a useful type of information which is not yet encoded in any version of PWN. The DOMAINS labels could be used in a similar way.

3.4. Annotation Transfer

Having a parallel corpus aligned at the word and phrase level may be the starting point on significant geographical and cross-lingual distribution of the tasks aimed at creating a multiple layers annotated corpus. One can imagine a cross-cultural initiative, which agreed on some parallel corpora containing the languages of interest, and where each partner is willing to annotate his/her language part of the multilingual corpora with the information for which the appropriate tools exists (POS, multi-word expressions, sense labels, parse tree annotations, argument/frame structures, etc). Based on the assumption that correcting annotations is easier and cheaper than creating them from scratch, word alignment technology could be used to transfer information from the tokens in one language to their translation equivalents in the other language. The transfer could be controlled by language specific rules the writing of which is certainly less demanding than the direct annotation.

An example at hand is transferring the word senses. Hand word sense disambiguation of a large text is an extremely labor intensive work, prone to human errors and extremely expensive.

SemCor2.0 is an English corpus, with (most of) the content words being sense disambiguated and carefully validated. It is not surprising that several research teams (see for instance http://multisemcor.itc.it/index.php) decided to translate as much SemCor documents as possible and then to transfer the senses into the translations. As we have shown in section 3.2, we partially translated the SemCor2.0 and used word-alignments to check out the completeness of our wordnet. We showed that almost 12,000 Romanian words translating the English sense annotated words were absent from our wordnet. By sense transfer, we can extend the target wordnet with almost 12,000 synsets. It is true that these synsets are partial (they are mono-literals) and would certainly require extensions with other synonyms but much effort is already saved. Similar considerations apply for the automatic extension of the 7000 incomplete synsets discovered by the experiment we described in section 3.2.

Another type of annotation transfer, more difficult and less reliable, but extremely useful, refers to the syntactic/semantic relations annotated in the source language. When the same type of syntactic/semantic annotation exists in both languages, the annotation transfer allows for annotation validation in one or both languages of the bitext, or provides evidence for corpus-supported comparative/contrastive studies.

In [1], [2] it is described an experiment aimed at assessing the possibility of statistically inducing a dependency grammar for Romanian by semi-automatic transfer of the dependency relations from a parsed

[6] concediat (Ro) = fired (En)

English text. The major assumption to be evaluated was the so-called Direct Correspondence Assumption (DCA) hypothesized in [11] for parallel treebanks. Put in terms of our source annotations, the DCA comes at checking whether a dependency relation that holds between two English words remains valid between the Romanian translation equivalents in the aligned bitext. In what follows we present the main lines of the strategy employed in this experiment and some preliminary results.

The aligned English-Romanian bitext used for this study was extracted from the *1984* parallel corpus; the English part of the bitext was parsed with the FDG parser [23] by a partner in Wolverhampton and validated by another partner in Iaşi.

From the entire bitext only 1537 sentence pairs (about 25%) were retained for the proper experiment and the sentence and word alignment were hand validated and corrected where necessary. We discarded very long sentences, the non-1:1 translation units, and those translation units with fewer aligned words than an empirical threshold or containing slang and non-grammatical language (*parole* language).

The relations between English words with a NULL translation equivalent were not taken into account for the evaluation of the transfer accuracy (in Table 2 below, these are counted in the **Lost** column), while the rest of relations were mechanically transferred into the Romanian part of the bitext. Two experts independently evaluated the validity of the transferred relations with disagreements negotiated and, agreed one way or another. There were identified three types of relation transfer: for the first type, the transfer is possible and correct without amendments; the second type refers to correct link transfer but incorrect labeling of the links; it needs mapping rules for switching the names of the correct link dependencies (e.g. the rule responsible for an active voice construction in English, translated by a passive voice construction in Romanian would switch the **obj** and **subj** labels in the target language sentence); the third category of transfers refers to the "lexicalized" dependencies (relations whose governor (rarely the dependent) is instantiated by a specific word) where the transfer is always wrong (both the dependency link and its name), due to the different behavior of corresponding predicates in the considered languages (e.g. like/plăcea).

Table 2. Percentage of correctly transferred relations

No.	Rel.	RO	Lost	EN	Acc. (%)
1	qn	10	0	12	83.33
2	neg	10	0	13	76.92
3	oc	3	0	4	75.00
4	dat	3	0	4	75.00
5	cnt	8	0	11	72.73
6	ad	25	0	35	71.43
7	pcomp	218	9	316	71.01
8	det	126	173	355	69.23
9	comp	70	1	112	63.06
10	attr	151	4	245	62.66

No.	Rel.	RO	Lost	EN	Acc (%)
11	Cc	94	2	155	61.44
12	Pm	44	1	75	59.46
13	Obj	79	2	137	58.52
14	Mod	114	1	201	57.00
15	Ha	41	0	74	55.41
16	Cla	8	0	15	53.33
17	Tmp	23	0	46	50.00
18	Man	16	0	32	50.00
19	Subj	121	72	319	48.99
20	v-ch	35	48	143	36.84

Table 2 gives information on the correctness of the unconditional transfer for the relations[7] in the source part of the bitext (for the description of these relations, as well as several examples, see [23]). Due to the enclitic definite articulation in Romanian, half of the English determiners (the occurrences of **the**) are not explicitly translated and consequently, half of the **det** relations are lost. The large number of the **subj** relations that are lost is due to the pro-drop nature of Romanian. One also may notice that this relation has a low correct transfer figure (48.99%) which is correlated with the low correct transfer figure of the relation **obj**. The simple mapping rule, mentioned before, for dealing with passive/active voice alternation in the aligned sentences would improve with almost 50% the success rate.

We consider these results as extremely encouraging, and one of our future research topics will be the design of a set of transfer rules for correcting the role assignment for the dependency links which were correctly transferred (second type, see above). The "lexicalized" dependencies will be collected as they would be detected and stored (with the correct transfer) as exceptions from the general transfer procedure.

A similar transfer experiment, but this time involving the valency frames existing in the Czech wordnet of the Balkanet project was carried on with Romanian wordnets as target. We used 601 valency frames, kindly offered by the Czech partner in Balkanet and the Czech-Romanian aligned sub-corpus of the "1984" parallel corpus. The manual validation of the automatic transfer of the Czech valency frames from the Czech

[7] The **phr** relation is not included being specific for English phrasal verbs.

verbs to their Romanian translation equivalents revealed a surprisingly high matching (80%), given the differences between Slavic and Romance languages.

4. COLLOCATION ANALYSIS IN A PARALLEL CORPUS

Once a parallel corpus has been word-aligned a very interesting cross-lingual study can be achieved in the area of multilingual terminology, multiword expressions and collocational patterns. Within the context of a trilateral project (University Marc Bloch from Strasbourg, IMS Stuttgart University and RACAI) we experimented on a large four-language parallel corpus (En-Ro-Fr-Ge) extracted from the Acquis Communautaire (AcqCom) multilingual corpus. For the word-alignment we used the English text as a hub language and after generating the En-Ro, En-Fr and En-Ge alignments, by transitivity we computed the alignments Ro-Fr, Ro-Ge and Fr-Ge. These last three alignments were combined, as discussed in Section 3, with the corresponding Ro-Fr, Ro-Ge and Fr-Ge alignments directly generated from the parallel corpus, thus obtaining more accurate alignments.

For the texts in each language in this parallel corpus the collocations were independently extracted (RACAI did it for Romanian and English, IMS for German and University Marc Bloch of Strasbourg for French). Our collocation extraction algorithm is similar to Smadja and McKeown's approach [22]. Based on the word alignment of the different bitexts, one could extract the translations in one language of the collocations detected in the other languages. At the time of the writing of this paper, we performed the partial analysis of the collocations in Romanian and English with respect to their translations in English and Romanian respectively[8]. We selected the best scored 20.000 independently extracted collocations in English ($COLLOC_{EN}$) and Romanian ($COLLOC_{RO}$). Then, by translation equivalence relations, found by the word aligner, we identified the translations into Romanian of the English collocations (TR_{RO}-$COLLOC_{EN}$) and the translations into English of the Romanian collocations (TR_{EN}-$COLLOC_{RO}$).

Given that the corpus contains specific uses and specialized language most of the collocations represent specific multi-word terms and most of them have word by word translation (Member State = Stat Membru, administrative transparency = transparență administrativă, act of accession = act de aderare, enter into force = intra în vigoare, etc.). The vast majority of these collocations were found in the intersection of the sets:

$$\text{SURE-COLLOC}_X = \text{COLLOC}_X \cap \text{TR}_X\text{-COLLOC}_Y \quad (1)$$

with X=English &Y=Romanian or X=Romanian &Y English respectively.
However, the most interesting collocations were those not found in the previous intersection sets:

$$\text{INTERESTING-COLLOC}_Z = \text{COLLOC}_Z \setminus \text{SURE-COLLOC}_Z \quad (2)$$

with Z=English or Romanian.

The multiword expressions in the lists INTERESTING-$COLLOC_Z$ were hand validated for termhood, cleaned-up and classified[9] into three major cases[10]:

a) aligner failure to detect the equivalence, due to preprocessing error and its imperfect RECALL (ex: "in vitro", "in vivo"; in Romanian these words were both wrongly tagged and lemmatized)
b) aligner failure to detect the equivalence due to a free human translation of the original text
c) aligner failure to detect the equivalence due to a non-word-by-word translation of the terms (especially those containing light-verbs).

An example of the case b) is given by the following original English text:

"*Whereas under Article 6 of the abovementioned Regulation the time when a transaction is carried out is considered as being the date on which occurs* ***event****, as defined by Community rules or, in the absence of*

[8] A similar study is conducted by our project partners in Strasbourg and Stuttgart but we don't have their results yet.

[9] We provide only qualitative evaluation as the results are still under investigation. A thorough analysis report for all the language pairs of the project is expected by March 2007.

[10] We ignore here pathological cases such as wrong human translation, segments of text not translated, texts in other languages than the expected ones, etc. These situations were quite numerous, and most of them were detected by our processing tools, but the monolingually extracted collocations obviously differ from the collocations obtained by translation equivalence.

and pending adoption of such rules, by the rules of the Member State concerned, ***in which the amount*** *involved in the transaction* ***becomes due and payable****.*", which was translated as:

"*Întrucât, conform art. 6 din regulamentul menționat anterior, se consideră ca moment al realizării operației data la care intervine* ***faptul generator de creanță*** *legată de valoarea aferentă operației respective, așa cum este el definit de reglementarea comunitară sau, dacă aceasta nu există și urmează a fi adoptată, de reglementarea statului membru interesat.*"

In this example the scattered English text: "*(the) event... in which the amount... becomes due and payable*" corresponds to the Romanian term "*situație generatoare de creanțe*" (a literal translation would be "(a) situation generating dues")

The last category is the most interesting as it outlines the multiword expressions which, due to their structural differences, are the hardest to translate by a simple-minded word-by-word approach. They range from legalese jargon (e.g. adversely affect <-> *a aduce atingere*[11]; legal remedy <-> *cale de atac*[12], to make good the damage <-> *a compensa daunele*[13] etc.) to constructions which are language and culture specific: for instance in Romanian somebody *takes* (and not makes) *a decision*, but *makes* (not take) *a shower*, something *makes* (not is) *part of*, etc.). Failing to use the exact wording of such a multiword expression, usually, is the major error source for language comprehension/production by language learners, as well as for other human beings in need to communicate but constrained to use a foreign language.

An interesting preliminary contrastive report on the light-verbs based collocations, with a case study of the verb *a face* (to do/make) for the French-Romanian AcqCom data is presented in [24].

We plan to develop a multilingual collocation dictionary, placing the major emphasis on the "hard" collocations (those existing in at least one inventory INTERESTING-COLLOC$_Z$ with Z one of the project languages) providing structural descriptions, translations in all considered languages, morpholexical restrictions on constituents (such as obligatory definiteness/indefiniteness, singular/plural, obligatory case, etc). We aim at a unified description of the collocational patterns in the four languages (with a perspective to extend our work to all the languages represented in the Acquis Communautaire corpus) and the development of a comprehensive multilingual dictionary, essential for dealing with the hard topic of collocation translation.

5. WEB SERVICES

We implemented a NLP web-services platform that currently ensures the basic preprocessing steps (tokenization-including multiword expression recognition, tiered tagging, lemmatization, language identification, sentence alignment) for English and Romanian corpora, as well as a search engine for the English-Romanian parallel corpora[14].

The services are implemented using standard technology (SOAP/WSDL/UDDI) on a dedicated bi-processor server with a reasonable high-speed Internet connection (100Mb/s). The NLP web-services will be continuously extended with new services (word alignment, collocation extraction, aligned-wordnets search, translation, QA in open domains, summarization, mono- and multilingual word sense disambiguation, etc.). Although most of the present (and near-future) services are available only for Romanian and English, we plan to add as many new languages as possible. The CLARIN initiative (http://www.clarin.eu/), recently included into the European Roadmap for Research Infrastructures[15], has been adhered by 47 institutions from 30 European countries and it is supposed to be the major collaborative work environment that will create, adapt and maintain the language resources and tools we will add to our NLP web services platform.

The current web-services were already used by various remote project partners in the LT4L European project (http://www.lt4el.eu/) and in the cross-lingual QA evaluation exercise at CLEF 2006 (http://www.clef-campaign.org/2006/ working_notes/CLEF2006WN-Contents.html).

[11] A mot-a-mot translation would be *to bring a touch*

[12] A mot-a-mot translation would be *way to attack*

[13] A mot-a-mot translation would be *to compensate the damages*

[14] Due to the upgrading of the web server and temporary security reasons, the services are available only to a few close collaborating teams in Romania, Germany and France, but soon the access will be open to the general community on a GNU-type of license.

[15] http://cordis.europa.eu/esfri/roadmap.htm

6. CONCLUSIONS

Aligned parallel corpora are of crucial importance to the multilingual processing. For cross-linguistic studies and validations, annotation transfer or for multilingual text processing, these texts represent the starting point of any idea of multilingual processing and resources.

In this paper we have seen that aligned parallel texts can be used for the benefit of a variety of applications that include:

- semantic validation of conceptually-aligned semantic lexical networks;
- semantic and syntactic annotation transfer;
- multilingual word sense disambiguation;
- cross-lingual collocation discovery.

Besides those reported here, aligned parallel corpora are useful in applications such as cross-lingual information retrieval and question answering, machine translation, multilingual dictionary extraction, and so forth.

The recent advances in NLP technology, demonstrate the tremendous benefits of collecting and adequately encoding large parallel corpora and multilingual semantic lexicons and ontologies. Building, in a concerted way, this kind of resources, for as many languages and as large as possible, should be a constant objective for an internationally established research infrastructure. Initiatives like DAM-LR (www.mpi.nl/DAM-LR/home.html), Global WordNet (http://www.globalwordnet.org/), Language Grid (http://langrid.nict.go.jp/), the previously mentioned CLARIN initiative and a few others are pioneering this ever-increasing need.

REFERENCES

1. BARBU-MITITELU V., ION R., *Cross-language Transfer of Syntactic Relations Using Parallel Corpora*. Proceedings of the Workshop on Cross-Language Knowledge Induction, EUROLAN'2005, Cluj-Napoca, Romania, pp. 46-51, 2005.
2. BARBU-MITITELU V., ION R., *Automatic Import of Verbal Syntactic Relations Using Parallel Corpora*. In Proceedings of the International Conference Recent Advances in Natural Language Processing, Borovets, Bulgaria, 21-23 septembrie 2005, pp. 329-332, 2005.
3. BRANTS T., *TnT a statistical part-of-speech tagger*. Proceedings of the 6th ANLP Conference, Seattle, WA., pp. 224-231, 2000.
4. BROWN P. F., DELLA PIETRA S.A., DELLA PIETRA V. J., MERCER R. L., *The mathematics of statistical machine translation: Parameter estimation*. Computational Linguistics, **19**(2) pp. 263–311.
5. CEAUŞU A., *Maximum Entropy Tiered Tagging*. Proceedings of the Eleventh ESSLLI Student Session, Malaga, Spain, pp. 173-179, 2006.
6. CEAUŞU A., ŞTEFĂNESCU D., TUFIŞ D., *Acquis Communautaire sentence alignment using Support Vector Machines*. Proceedings of the 5th LREC Conference, Genoa, Italy, pp. 2134-2137, 2006.
7. ERJAVEC T., *MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora*. Proceedings of the 4th LREC Conference, Lisbon, Portugal, pp. 1535 – 1538, 2004.
8. FAN R.-E., CHEN P.-H., LIN C.-J., *Working set selection using the second order information for training SVM*. Technical report, Department of Computer Science, National Taiwan University, www.csie.ntu.edu.tw/ ~cjlin/papers/quadworkset.pdf, 2005.
9. FELLBAUM C. (ed.), *WordNet: An Electronic Lexical Database*, MIT Press, 1998.
10. GALE W.A., CHURCH K.W., *A Program for Aligning Sentences in Bilingual Corpora*. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California, USA, pp. 177–184, 1991.
11. HWA R., RESNIK P., WEINBERG A., KOLAK O., *Evaluating translational correspondence using annotation projection*. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, 2002.
12. ION R., *Metode de dezambiguizare semantică automată. Aplicații pentru limbile engleză și română*. PhD Thesis, Romanian Academy, Bucharest , Romania, 145 p.
13. ION R., BARBU-MITITELU V., *Constrained Lexical Attraction Models*. In Proceedings of FLAIRS2006, Melbourne Beach, Florida, USA., 2006.
14. ION R., CEAUŞU A., TUFIŞ D., *Dependency-Based Phrase Alignment*. In Proceedings of the 5th LREC Conference, Genoa, Italy, pp. 1290-1293, 2006.
15. ION R., TUFIŞ D., *Multilingual Word Sense Disambiguation Using Aligned Wordnets*. Romanian Journal on Information Science and Technology, Tufiş D. (ed.) Special Issue on BalkaNet, Romanian Academy, **vol 7**, no. 2-3, pp. 198-214, 2004.
16. MAGNINI B., CAVAGLIÀ G., *Integrating Subject Field Codes into WordNet*. In Proceedings of LREC2000, Athens, Greece, pp. 1413-1418, 2000.
17. MARTIN J., MIHALCEA R., PEDERSEN T., *Word Alignment for Languages with Scarce Resources*. In Proceeding of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond". June, 2005, Ann Arbor, Michigan, June, Association for Computational Linguistics, 65–74, 2005.

18. MOORE R. C., *Fast and Accurate Sentence Alignment of Bilingual Corpora*. In Machine Translation: From Research to Real Users. Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California, Springer-Verlag, Heidelberg, Germany, pp. 135-244, 2002.
19. NILES I., PEASE A., *Towards a Standard Upper Ontology*. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Ogunquit, Maine, pp. 17-19, 2001.
20. OCH F.J., NEY H., *Improved Statistical Alignment Models*, In Proceedings of ACL2000, Hong Kong, China, pp. 440-447, 2000.
21. OCH F.J., NEY H., *A Systematic Comparison of Various Statistical Alignment Models*, Computational Linguistics, **29**(1), pp. 19-51, 2003.
22. SMADJA F. A., McKEOWN K. R., *Automatically extracting and representing collocations for language generation*, Proceedings of the 28th annual meeting on Association for Computational Linguistics, pp. 252-259, Pittsburgh, Pennsylvania, 1990.
23. TAPANAINEN P., JÄRVINEN T., *A dependency parser for English*. Technical Report no. TR-1, Department of General Linguistics, University of Helsinki, Finland, 1997.
24. TODIRAŞCU A., *Towards an automatic extraction of collocations; the case of the verb MAKE/DO (in Romanian)*. In Proceedings of the National Workshop on Romanian Language Processing, Iaşi, 3-4 November, 2006.
25. TUFIŞ D., BARBU A., ION R., *Extracting Multilingual Lexicons from Parallel Corpora*, Computers and the Humanities, **Vol. 38**, Issue 2, May 2004, pp. 163 – 189, 2004.
26. TUFIŞ D., ION R., CEAUŞU A., ŞTEFĂNESCU D., *Improved Lexical Alignment by Combining Multiple Reified Alignments*. In Proceedings of the 11th Conference of the European Association for Computational Linguistics (EACL), Trento, Italy, April, pp. 153-160, 2006.
27. TUFIŞ D., ION R., IDE N., *Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets*. In proceedings of the 20th COLING Conference, Geneva, pp. 1312-1318, 2004.
28. TUFIŞ D., ION R. CEAUŞU A., ŞTEFĂNESCU D., *Combined Aligners*. In Proceedings of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond". Ann Arbor, Michigan, June, Association for Computational Linguistics, pp. 107-110, 2005.
29. TUFIŞ D., ION R., *Evaluating the word sense disambiguation accuracy with three different sense inventories*. In Proceedings of the Natural Language Understanding and Cognitive Systems Symposium, Miami, Florida, May, pp. 118-127, 2005.
30. TUFIŞ D., *Tiered Tagging and Combined Classifiers*. In F. Jelinek, E. Nöth (eds) Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin Heidelberg New-York, pp. 28-33, 1999.
31. TUFIŞ D., IDE N., ERJAVEC T., *Standardised Specifications, Development and Assessment of Large Morpho-Lexical Resources for Six Central and Eastern European Languages*. Proceedings of the 1st LREC Conference, Granada, Spain, pp. 233-240, 1998.
32. TUFIŞ D., CRISTEA D., STAMOU S., *BalkaNet: Aims, Methods, Results and Perspectives. A General Overview*. In Romanian Journal on Information Science and Technology, Tufiş D.(ed.) Special Issue on BalkaNet, Romanian Academy, **vol. 7**, no. 2-3, pp. 9-34, 2004.
33. TUFIŞ D., BARBU-MITITELU V., BOZIANU L., MIHĂILĂ C., *Romanian WordNet: New Developments and Applications*. Proceedings of the 3rd Conference of the Global WordNet Association, Jeju, Republic of Korea, pp. 337-344, 2006.
34. TUFIŞ D., ION R., IRIMIA E., BARBU-MITITELU V., CEAUŞU A., ŞTEFĂNESCU D., BOZIANU L., MIHĂILĂ C., *Resources, Tools and Algorithms for the Semantic Web*. Proceedings of the Workshop "IST – Multidisciplinary approaches", Bucharest, Romania, pp. 13-22, 2006.
35. VOSSEN P. (ed.), *A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht, 1998.