REIFYING THE ALIGNMENTS

Dan TUFIŞ, Radu ION, Alexandru CEAUŞU,
and Dan ŞTEFĂNESCU

Research Institute for Artificial Intelligence, Romanian Academy
E-mail: {tufis,radu,aceausu,danstef}@racai.ro

Abstract. Reification is the process of dealing with abstract entities as if they have a concrete or material existence. A bitext word alignment is a set of links between textual units (phrases or words) which are reciprocal translations. We associate to every link of an alignment a complex object, represented by a feature-value structure containing information about the linked lexical tokens (the translation probability, part-of-speech affinity, orthographic similarity etc.). Based on this structure, one link of an alignment can be evaluated as correct or not on an individual basis. We describe a procedure for word alignment of parallel texts, included into a larger platform that ensures the bitext pre- and post-processing (sentence alignment, tokenization, POS-tagging, lemmatization, chunking, XML encoding). The alignment procedure combines the results produced by two (or more) different word-aligners. We describe the base word aligners in some details and their individual evaluations. The combined aligner takes the union of the individual alignments as input and, by a SVM-based classification engine, filters out the improbable links. The evaluation shows that the combined word alignment has a 12.4% improved Alignment Error Rate (AER) *versus* the best individual aligner.

Key words: word alignment, word alignment combination, word alignment reification.

1. INTRODUCTION

The alignment process is a fundamental operation in building translation models which represent the backbone of any statistical machine translation (SMT) system. Therefore, alignment has been, beginning with early 90's (Brown *et al*., 1993), (Kay & Röscheisen, 1993), (Ahrenberg *et al*., 1998) etc., and continues to be a major preoccupation for the SMT community. The great impact of the public releases of GIZA (Al-Onaizan *et al*., 1999), GIZA++ (Och & Ney, 2000), the SMT developments kits such MOSES and its recent powerful enhancements (Koehn *et al*., 2007), word alignment shared tracks organized by NAACL2003 (Mihalcea &

Pedersen, 2003) or ACL2005 (Martin *et al*., 2005) and the vast literature in the last decade are just a few facts supporting the relevance of the word alignment process within a SMT system. Recent developments are characterized by relying more and more on linguistic features (lemmas, part-of-speech, parse trees) in their underlying methods, new statistical models and training strategies.

A pair of texts that represent the translation of each other is called a parallel text or a *bitext*. Aligning the different textual units (paragraphs, sentences, phrases or words) of a bitext is a process based on the notion of *translation equivalence*. In a given parallel text, the assumption is that the *same* meaning is linguistically expressed in two or more languages. Meaning identity between two or more representations of presumably the same thing is a notorious philosophical problem and even in more precise contexts than language (for instance, in software engineering) it remains a fuzzy concept. Consequently the notion of translation equivalence relation, built on the meaning identity assumption, is inherently vague. In the area of machine translation, terminology, multilingual information retrieval and other related domains, one needs operational notions, defined in precise, quantifiable terms. One of the widely accepted interpretations (Melamed, 2001) of the translation equivalence defines it as a relation that holds between two fragments of different language texts such that expressions appearing in *corresponding parts* of the two texts are reciprocal translations. These expressions are called *translation equivalents*. A bitext with its translation equivalents linked is called an *aligned bitext*. The granularity at which translation equivalents are defined (paragraph, sentence, phrasal, lexical) specifies the *granularity* of a bitext alignment (paragraph, sentence, phrasal, lexical). A pair of textual units where the translation equivalents are looked for is usually called a *translation unit*. The granularity of translation units can be defined in a similar way to the granularity of translation equivalents with the obvious difference that the textual span of a translation unit is larger than that of a translation equivalent. In this paper we will address the finest granularity level of a bitext alignment, namely the lexical alignment, most frequently referred to as *word alignment*. Most often than not, the translation unit for the word alignment is considered to be a p*air of corresponding* sentences or paragraphs from a bitext. The identification of pairs of words $<w^i_{L1}\ w^j_{L2}>$ that represent mutual translations 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 was 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*.

Many of the modern approaches to lexical alignment rely on statistical techniques and they roughly fall into two categories. The *hypotheses-testing* methods such as (Gale & Church, 1991), (Smadja *et al*., 1996) etc. use a hypotheses generator that produces a list of translation equivalence candidates (TECs), each of them being subject to an independence statistical test. The TECs

that show an association measure higher than expected under the independence assumption are assumed to be translation-equivalence pairs (TEPs). The TEPs are extracted independently one of another and therefore the process might be characterised as a local maximisation (greedy) one. The *estimating* approaches (Brown *et al*., 1993), (Kupiec, 1993), (Hiemstra, 1997) etc. are based on building from data a statistical bitext model, the parameters of which are to be estimated according to a given set of assumptions. The bitext model allows for global maximisation of the translation equivalence relation, considering not individual translation equivalents but sets of translation equivalents (sometimes called *assignments*). There are pros and cons for each type of approach, some of them discussed in (Hiemstra, 1997).

Departing from the usual approach of using GIZA++, we developed two quite different word aligners, driven by two distinct objectives: the first one was motivated by a project aiming at the development of an interlingually aligned set of wordnets while the other one was developed within an SMT ongoing project. The first one was used for validating, against a multilingual corpus, the interlingual synset equivalences and also for word sense disambiguation (WSD) experiments. Although, initially, it was concerned only with open class words recorded in a wordnet, turning it into an "all words" aligner was not a difficult task. This word aligner, called **YAWA,** is a typical hypotheses testing implementation and is described in section 3.1.

A quite different solution (closer to the model estimation approach) from the one used by YAWA, is implemented in our second word aligner, called **MEBA**, described in section 3.2. It is a multiple parameter and multiple step algorithm using relevance thresholds specific to each parameter, but different from each step to the other. The implementation of MEBA was strongly influenced by the IBM models described in (Brown *et al*., 1993). We used GIZA++ (Och & Ney, 2000; Och & Ney, 2003) to estimate some parameters of the MEBA aligner.

Both aligners use several features to characterize the links of an alignment. The main important link features are described in section 3.2.

The alignments produced by MEBA were compared to the ones produced by YAWA. Given that the two aligners are based on quite different approaches and that their F-measures are comparable, it was quite a natural idea to combine their results and hope for an improved alignment. Moreover, by analyzing the alignment errors done by each word aligner, we found that the number of common mistakes was small, so the premises for a successful combination were very good (Dietterich, 1998).

The Combined Word Aligner, **COWAL-**described in section 4, is a wrapper of the two aligners (YAWA and MEBA) merging the individual alignments and filtering the result. At the Shared Task on Word Alignment organized by the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond" (Martin *et al*., 2005), we participated (on the

Romanian-English track) with the two aligners and the combined one (COWAL). Out of 37 competing systems, COWAL was rated the first, MEBA the 20th and TREQ-AL, the former version of YAWA, was rated the 21st. This was a very convincing demonstration of the usefulness of the alignment combining approach.

Meanwhile, both the individual aligners and their combination were significantly improved. COWAL is more principled based and is now embedded into a larger platform (see Figure 1) that incorporates several tools for bitexts pre-processing (briefly reviewed in section 2), a graphical interface that allows for comparing and editing different alignments, as well as a word sense disambiguation module (not discussed here).

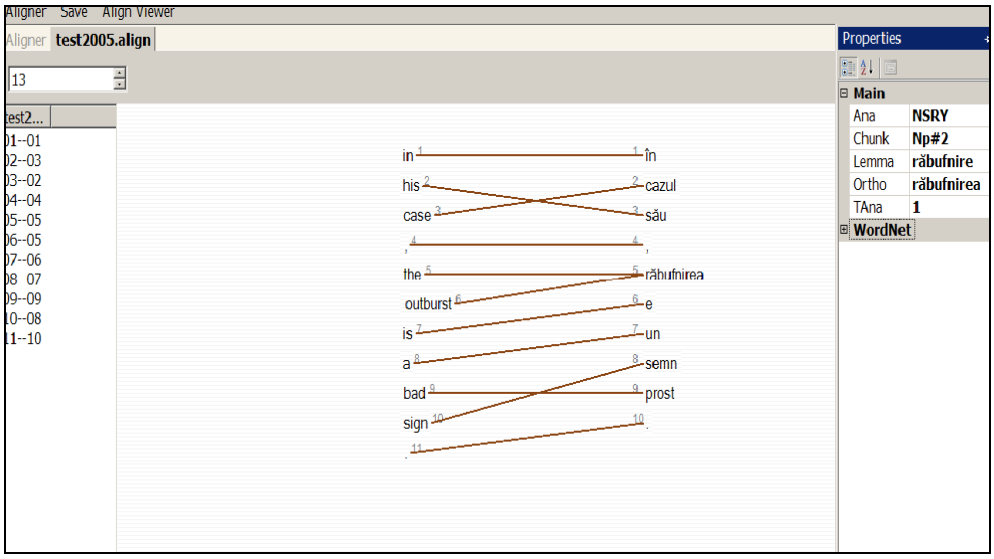

Figure 1. The alignment editor.

In the rest of the paper we will describe the present versions of the lexical aligners and evaluate their results. The evaluations were performed, using the official evaluation script, against the Gold Standards of the ACL2005 shared track (GS2005) and also against two recently corrected versions of the GS2005 (see section 4.2). For training, we used the data provided by the ACL2005 shared track, (which for English-Romanian pair of languages was the same as in NAACL2003 shared track). Additionally, we used the Gold Standard alignment from the 2003 competition.

The training data is a medium-sized English-Romanian corpus, containing approximately one million tokens per language. This corpus, compiled by Rada Mihalcea (http://www.cs.unt.edu/~rada/wpt/index.html#resources), groups together the parallel text of Orwell's novel "1984", the Romanian Constitution, and a large (about 800,000 tokens) collection of journalistic texts collected from the Web.

2. THE BITEXT PROCESSING

The two base aligners and their combination use the same format for the input data and provide the alignments in the same format. The input format is obtained from two raw texts that represent reciprocal translations. If not already sentence aligned, the two texts are aligned by our SVM sentence aligner (Ceauşu, Ştefănescu, Tufiş, 2006). The texts in each language are then tokenized, tagged and lemmatized by the TTL module (Ion, 2007).

Our word alignment algorithms, MEBA and YAWA, described in Section 3, require the following preprocessing steps 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 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 (Dimitrova *et al.*, 1998) with corresponding resources for Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene. Our segmenter is a part of the Perl module called TTL (Ion, 2007) and is able to recognize paragraphs, sentence and clause boundaries, dates, numbers and various fixed phrases, 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.*** 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 quality 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 (Tufiş, 1999) which 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 more 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 TTL (an HMM tagger), which produces also the lemmatization, and a maximum-entropy tagsets converter (Ceauşu, 2006) trained on texts manually tagged with both tagsets. 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 (Ion, 2007) a straightforward process, since the monolingual lexicons developed within MULTEXT-EAST contain, for each word, its lemma and the morpho-lexical tag. Currently this lexicon contains more than 1.3 million entries. 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.
- ***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 (Ion, 2007).

Finally, the bitext is assembled as an XML document which is the standard input for most of our tools, including COWAL alignment platform. All the mentioned pre-processing steps have been implemented as web services (Tufiş *et al*., 2008) using the SOAP/WSDL technology and recently have also been published as REST services on the WebLicht web-services platform developed within the CLARIN project (http://www.clarin.eu/).

3. THE BASE ALIGNERS

Although here we will consider only two base aligners as providers of reified alignments to be combined, there is no limitation on the number of alignments to be combined.

3.1. YAWA

YAWA is a three stage lexical aligner that uses bilingual translation lexicons and phrase boundaries detection to align words of a given bitext. The translation lexicons are generated by a different module (Tufiş, 2002), which produces

translation equivalence hypotheses for the pairs of words (one for each language in the parallel corpus) which have been observed occurring in aligned sentences more than expected by chance. The hypotheses are filtered by a log-likelihood score threshold (Dunning, 1993). Several heuristics (string similarity-cognates, POS affinities and alignments locality[1]) are used in a competitive linking manner (Melamed, 2001) to extract the most likely translation equivalents.

YAWA generates a bitext alignment by incrementally adding new links to those created at the end of the previous stage. The existing links act as contextual restrictors for the newly added links. From one phase to the other new links are added without deleting anything. This monotonic process requires a very high precision (at the price of a modest recall) for the first step. The next two steps are responsible for significantly improving the recall and ensuring an increased F-measure.

In the rest of this section we present the three stages of YAWA and evaluate the contribution of each of them to the final result.

Phase 1: Content Words Alignment

YAWA begins the alignment process by taking into account only very probable links that represent the skeleton alignment used by the second phase. This alignment is done using outside resources such as translation lexicons and involves only the alignment of content words (nouns, verbs, adjective and adverbs).

The translation equivalence pairs are ranked according to an association score (i.e. log-likelihood, DICE, point-wise mutual information, etc.). We found that the best filtering of the translation equivalents was the one based on the log-likelihood (*LL*) score with a threshold of 9. If T_T and T_S are target and source tokens, then the log-likelihood score is computed according to the formula:

$$LL(T_T, T_S) = 2 * \sum_{j=1}^{2} \sum_{i=1}^{2} n_{ij} * log \frac{n_{ij} * n_{**}}{n_{i*} * n_{*j}},$$

where:

- n_{11} = the number of parallel sentence pairs (PSP) in which T_S T_T co-occur;
- n_{12} = the number of PSP in which appeared T_S but not T_T;
- n_{21} = the number of PSP in which the T_T appeared but not T_S;
- n_{22} = the number of PSP in which neither T_S nor T_T appeared;
- n_{1*} = the number of PSP in which T_S appeared (irrespective of T_T);
- n_{*1} = the number of PSP in which T_T appeared (irrespective of T_S);
- n_{2*} = the number of PSP in which T_S did not appeared;

[1] The *alignments locality* heuristics exploits the observation made by several researchers that adjacent words of a text in the source language tend to align to adjacent words in the target language. A more strict alignment locality constraint requires that all alignment links starting from a chunk in the one language end in an aligned chunk of the other language.

- n_{*2} = the number of PSP in which T_T did not appeared;
- n_{**} = the total number of PSP.

Each translation unit (pair of aligned sentences) of the target bitext is scanned for establishing the most likely links based on a competitive linking strategy that takes into account the *LL* association scores given by the translation lexicon. If a candidate pair of words is not found in the translation lexicon, we compute their orthographic similarity (cognate score (Tufiş, 2002)). If this score is above a predetermined threshold (for Romanian-English bitexts we used the empirically found value of 0.43), the two words are treated as if they existed in the translation lexicon with a high association score (in practice we have multiplied the cognate score by 100 to yield scores in the range 0 … 100).

Phase 2: Chunks Alignment

The second phase requires that each part of the bitext is chunked. Currently, the chunking (for both Romanian and English) is driven by regular expressions defined over the tagsets used in the target bitext. These simple chunkers recognize noun phrases, prepositional phrases, verbal and adjectival or adverbial groupings in both languages.

YAWA produces first chunk-to-chunk matching and then aligns the words within the aligned chunks. Chunk alignment is done on the basis of the skeleton alignment produced in the first phase. The algorithm is simple: align two chunks $c(i)$ in source language and $c(j)$ in the target language if $c(i)$ and $c(j)$ have the same type (noun phrase, prepositional phrase, verb phrase, adjectival/adverbial phrase) and if there exist a link $\langle w(s), w(t)\rangle$ so that $w(s) \in c(i)$ then $w(t) \in c(j)$.

After alignment of the chunks, a language pair dependent module takes over to align the unaligned words belonging to the chunks. The mild language-pair dependency of YAWA is given by the requirement to customise (when necessary) a general heuristics which we refer to as *Head Linking Projection heuristics (HLP):*

if ***b*** *is aligned to* ***c*** *and* ***b*** *is preceded by* ***a*** *,then link* ***a*** *to* ***c*** *(case A in Figure 2)*

unless there exist ***d*** *in the same chunk with* ***c*** *and the POS category of* ***d*** *has a significant affinity with the category of* ***a*** *(case B in Figure 2).*

The simplicity of these rules derives from the shallow structures of the chunks. In the above rule ***b*** and ***c*** are content words while ***a*** is very likely a determiner or a modifier for ***b***. In Figure 2 (A and B) the heavy lines are links from the skeleton alignment (Phase 1) which by virtue of HLP induce new links (represented by the dash lines). This heuristics is sufficiently general to apply for a large number of language pairs.

Figure 2. Head Linking Projection Heuristics: bitexts are represented as vertical texts, side by side (***a*** and ***b*** belong to the first language; ***c*** and ***d*** belong to the second language).

Phase 3: Dealing with Sequences of Unaligned Words

This phase identifies contiguous sequences of words (blocks) in each part of the bitext which remained unlinked and heuristically attempts first to match them and them to link their words.

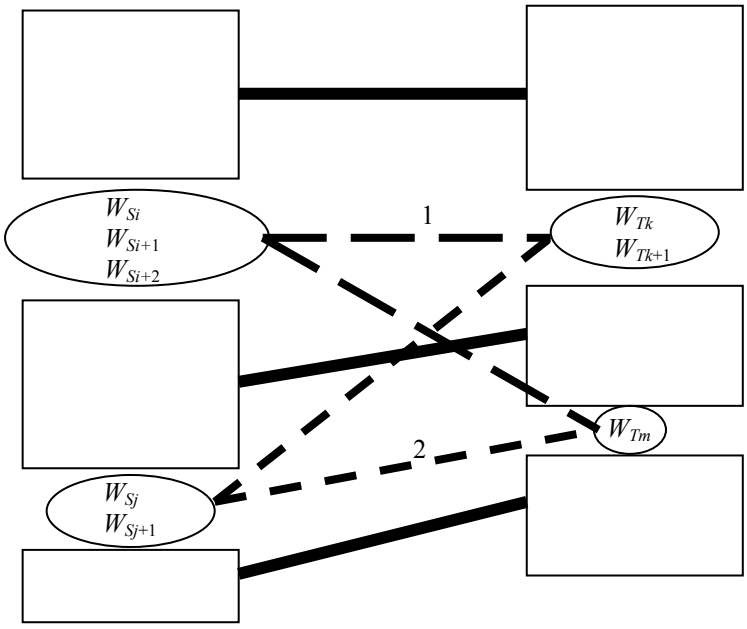

Figure 3. Aligning blocks of unaligned words.

In Figure 3 we represented by squares the chunks of texts in each side of the bitext which contains links created in the previous phases. Via the links they contain, these blocks are implicitly aligned (represented by heavy lines). The sequences of unlinked words in one part of the bitext are likely to be linked to similar sequences of unlinked words in the other part of the bitext. These sequences of words are circled in Figure 3 and the dashed lines represent possible block alignments. The major heuristics used in this step is that the sequences of words, external to the surrounding aligned segments of texts, could be linked to words belonging to

similar sequences in the other part of the bitext, situated in *plausible* positions. To detect which are the most plausible block alignments, several figures of merit are linearly interpolated: the ratio between the numbers of words in each candidate sequence, the number of the crossing with the already aligned blocks and the POS-affinities of the remaining unaligned words.

For the example in Figure 3, the *geometric criterion* (minimizing the number of crossing) would favour the alignments of the sequence (W_{Si}, W_{Si+1} W_{Si+2}) to the sequence (W_{Tk}, W_{Tk+1}) and of the sequence (W_{Sj}, W_{Sj+1}) to the sequence (W_{Tm}). These alignments marked in Figure 2 as 1 and 2 would not generate crossing links. The number of words criterion would favour the same sequence alignment because the other sequence alignment might produce a 3-1 alignment which is less probable than 2-1 or 1-1 alignment. Finally, the POS affinity between the words the candidate sequences contain represents the strongest heuristic criterion, since it is used to generate the initial links inside the aligned sequences of unaligned blocks. After block alignment, decided by maximization of the linear interpolation of the above three criteria, given a pair of aligned blocks, the algorithm links (in a competitive linking manner) the words with the same highest POS affinities and then the phase 2 is called again with these new links as the skeleton alignment.

The third phase is responsible for significant improvement of the alignment recall, but it also generates several wrong links. The detection of some of them is quite straightforward, and we added an additional correction phase 3.f. By analysing the bilingual training data we noticed the translators' tendency to preserve the order of the phrasal groups. We used this finding (which might not be valid for any language pair) as a removal heuristics for the links that cross two or more aligned phrase groups.

YAWA performance analysis

Table 1 presents the (cumulative) results of the YAWA aligner at the end of each alignment phase. Although the Precision decreases from one phase to the next one, the Recall gains are significantly higher, so the F-measure is monotonically increasing. The evaluation was performed against the most recent Gold Standard (GS3 – see section 4.2).

Table 1. YAWA evaluation

	Precision	**Recall**	**F-Measure**	**AER**
Phase 1	94.08%	34.99%	51.00%	49.00%
Phase 1+2	89.90%	53.90%	67.40%	32.60%
Phase 1+2+3	88.82%	73.44%	80.40%	19.60%
Phase 1+2+3+3.f	**88.80%**	**74.83%**	**81.22%**	**18.78%**

3.2. MEBA

MEBA uses an iterative algorithm that takes advantage of all pre-processing phases mentioned in section 2. Similar to YAWA aligner, MEBA generates the links step by step, beginning with the most probable (*anchor links*). The links to be added at any later step are supported or restricted by the links created in the previous iterations. The aligner has different weights and different significance thresholds on each feature and iteration. Each of the iterations can be configured to align different categories of tokens (named entities, dates and numbers, content words, functional words, punctuation) in decreasing order of statistical evidence and using the aligned pairs from previous steps as support for the new ones.

MEBA starts with building the so-called "anchoring" links (high confidence): tokens with *high cognate scores, identical POS, high translation equivalence scores, high translation equivalence entropy score (1-normalised entropy), or low obliqueness*. Then, it incrementally tries to link the words around the already linked tokens. The locality with respect to the anchoring links provides strong clues for adding new links. The candidate aligned pair that cross too many links already established are discouraged. Punctuation linking is addressed in a distinct step and the locality scores and the number of crossed links guide it. Finally, as in case of YAWA, the Head Feature Projection heuristics is used to link all the remaining unlinked words.

A link between two tokens is characterized by a set of features (with values in the [0, 1] interval). We differentiate between *context independent features* that refer only to the tokens of the current link (translation equivalency, part-of-speech affinity, cognates, etc.) and *context dependent features* that refer to the properties of the current link with respect to the rest of links in a bitext (locality, number of traversed links, tokens' index displacement and collocation). Also, we distinguish between bidirectional features for which the values are computed in both directions (translation equivalence, translation equivalence entropy, part-of-speech affinity) and non-directional features (cognates, locality, number of traversed links, collocation, obliqueness-relative positions of the linked tokens).

The score of a candidate link (*LS*) between a source token *i* and a target token *j* is computed by a linear function of several features scores (Tiedemann, 2003).

$$LS(i,j) = \sum_{i=1}^{n} \lambda_i * ScoreFeat_i \; ; \sum_{i=1}^{n} \lambda_i = 1$$

Each feature has defined a specific significance threshold, and if the feature's value is below this threshold, the contribution to the *LS* of the current link of the feature in case is 0.

The thresholds of the features and lambdas are different from one iteration to the next one and they are set by the user during the training and system fine-tuning phases.

There is also a general threshold for the link scores and only the links that have the LS above this threshold are retained in the bitext alignment. Given that this condition is not imposing unique source or target indexes, the resulting alignment is inherently many-to-many.

Table 2 summarizes the (cumulative) results after each of the main iterations mentioned earlier. The evaluation was performed against the most recent Gold Standard (GS3 – see section 4.2).

Table 2. MEBA evaluation.

	Precision	**Recall**	**F-Measure**	**AER**
"anchor" words	98.06%	29.00%	44.76%	55.24%
words around "anchors"	96.39%	43.53%	59.98%	40.02%
minimally crossing alignments	95.38%	50.58%	66.36%	33.64%
punctuation	94.71%	60.08%	73.52%	26.64%
head linking projection	92.05%	71.00%	80.17%	19.83%

The features describing a link

In the following subsections we briefly discuss the main features MEBA uses in characterising a link. Some of these features are used by YAWA too (translation equivalence scores, cognate, POS affinity, locality and obliqueness).

Translation equivalence

This feature (TE) may be used for two types of pre-processed data: lemmatized or non-lemmatized input. Depending on the input format, MEBA invokes GIZA++ to build translation probability lists for either lemmas or the occurrence forms of the bitext[2]. Irrespective of the lemmatisation option, the considered token for the translation model build by GIZA++ is the respective lexical item (lemma or word-form) trailed by its POS tag (eg. plane_N, plane_V, plane_A). In this way we avoid data sparseness and filter noisy data. For instance, in case of highly inflectional languages (as Romanian is) the use of lemmas significantly reduces the data sparseness. For languages with weak inflectional character (as English is) the POS trailing contributes especially to the filtering the search space. A further way of removing the noise created by GIZA++ is to filter out all the translation pairs with a log-likelihood (*LL*) score below a predefined threshold.

As in the case of YAWA, we made various experiments and, based on the estimated ratio between the number of false negatives and false positive, empirically set the same value of this threshold (9). All the probability losses by

[2] Actually, this is a user-set parameter of the MEBA aligner; if the input bitext contains lemmatization information, both translation probability tables may be requested.

this filtering were redistributed proportionally to the initial probabilities to the surviving translation equivalence candidates.

Translation equivalence entropy score

The translation equivalence relation is a semantic one and it directly addresses the notion of word sense. In a coherent text, the distribution of the senses of a word occurring several times is expected to be a skewed one (Zipf, 1936). We used this conjecture as a highly informative information source for the validity of a candidate link. The translation equivalence entropy score is a favouring parameter for the words that have unevenly distributed translation probabilities.

$$ES(W) = 1 + \frac{\sum_{i=1}^{N} p(TR_i|W) * \log p(TR_i|W)}{\log N}$$

where: W is the token for which the entropy score is computed;
TR_i is one of the possible translations of the token W;
N is the number of the possible translations;
and $p(TR_i|W)$ is the estimated probability of the translation TR_i for the token W.

Part-of-speech affinity

Melamed (1996) observed that more often than not, the translation equivalents have the same part-of-speech, that is, most of the time a verb translates as a verb, a noun as a noun and so on. He called such translation pairs V-type, to distinguish them from those translation pairs where the part of speech of one token in not the same as the one for the other token. This type was called P-type translation pairs. A word aligner producing only V-type links would presumably have a high precision but its recall would seriously be affected. When the translation equivalents have different parts of speech, this difference is not arbitrary, but restricted by what we called *POS-affinity*. The POS-affinity probabilities (PA for short) are conditional probabilities that can be easily estimated from an already aligned tagged bitext (we used the trial data and the GS2003):

$$\mathrm{PA}\left(\mathrm{POS}_m^{\mathrm{Source}} \middle| \mathrm{POS}_n^{\mathrm{Target}}\right) = \frac{\#\mathrm{links}\left\langle \mathrm{POS}_m^{\mathrm{Source}}, \mathrm{POS}_n^{\mathrm{Target}} \right\rangle}{\#\mathrm{links}\left\langle \mathrm{POS}_m^{\mathrm{Source}}, \mathrm{anyPOS}^{\mathrm{Target}} \right\rangle}$$

$$\mathrm{PA}\left(\mathrm{POS}_n^{\mathrm{Target}} \middle| \mathrm{POS}_m^{\mathrm{Source}}\right) = \frac{\#\mathrm{links}\left\langle \mathrm{POS}_m^{\mathrm{Target}}, \mathrm{POS}_n^{\mathrm{Source}} \right\rangle}{\#\mathrm{links}\left\langle \mathrm{POS}_m^{\mathrm{Target}}, \mathrm{anyPOS}^{\mathrm{Source}} \right\rangle}$$

Cognates

The similarity measure we used, SYM (T_S, T_T), is very similar to the **XXDICE** score described in (Brew & McKelvie, 1996). If T_S is a string of k characters $\alpha_1\alpha_2 \ldots \alpha_k$ and T_T is a string of m characters $\beta_1\beta_2 \ldots \beta_m$ then we construct two new strings T'_S and T'_T by inserting where necessary special displacement characters into T_S and T_T. The displacement characters will cause both T'_S and T'_T have the same length p ($max\ (k, m) \leq p < k+m$) and a maximum number of positional matches. Let $\delta(\alpha_i)$ be the number of displacement characters that immediately precede the character α_i which matches the character β_i and $\delta(\beta_i)$ be the number of displacement characters that immediately precede the character β_i which matches the character α_i. Let q be the number of matching characters. With these notations, the SYM (T_S, T_T) measure and the $COGN$ feature value are defined as follows:

$$SYM(T_S, T_T) = \begin{cases} \sum_{i=1}^{q} \frac{2}{1+|\delta(\alpha_i)-\delta(\beta_i)|}, & \text{if } q > 2 \\ \frac{}{k+m} \\ 0, & \text{if } q \leq 2 \end{cases}$$

$$COGN(T_S, T_T) = \begin{cases} 1, & \text{if } SYM(T_S, T_T) > \text{Threshold} \\ 0, & \text{otherwise} \end{cases}$$

The threshold for the $SYM(T_S, T_T)$ was empirically set to 0.42. This value depends on the pair of languages in the considered bitext. The actual implementation of the *SYM* function considers a language dependent normalisation step, which strips some suffixes, discards the diacritics and reduces some consonant doubling etc. This normalisation step was hand written, but, based on available lists of cognates, it could be automatically induced. Unlike MEBA which uses the *COGN* feature, YAWA uses *SYM* score. It is interesting to notice that the *COGN* feature is relevant for most pairs of languages, although for languages with different scripts an additional transliteration step would be necessary.

Obliqueness

Each token in both sides of a bi-text is characterized by a position index, computed as the ratio between the relative position in the sentence and the length of the sentence. The absolute value of the difference between tokens' position indexes, subtracted from 1[3], gives the link's "obliqueness". This feature is "context free" as opposed to the locality feature described in the next sub-section.

[3] This is to ensure (as in the case of the *ES* score) that values close to 1 are "good" ones and those near 0 are "bad". This definition takes into account the relatively similar word order in English and Romanian.

$$OBL(SW_i, TW_j) = 1 - \left| \frac{i}{length(Sent_S)} - \frac{j}{length(Sent_T)} \right|$$

where: *SW* is the source token;
TW is the target token;
i is the index of the source token in the source sentence ($Sent_S$);
and *j* is the index of the source token in the target sentence ($Sent_T$).

Locality

Locality is a feature that estimates the degree to which the links are sticking together. MEBA has two features to account for locality: (i) *weak locality* and (ii) *chunk-based locality*. The value of the *weak locality* feature is derived from the already existing alignments in a window of *N* links centred on the candidate new link $<S_\alpha\ T_\alpha>$, see Figure 4.

The window size is variable, proportional to the sentence length. If in the window there exist k linked tokens and the relative positions of the tokens in these links are $<i_1\ j_1>$, ...$<i_k\ j_k>$ then the locality feature of the new link $<i_{k+1}, j_{k+1}>$ is defined by the equation below:

$$LOC = \frac{1}{k} \sum_{m=1}^{k} \frac{\min(|\, s_\alpha - s_m \,|, |\, t_\alpha - t_m \,|)}{\max(|\, s_\alpha - s_m \,|, |\, t_\alpha - t_m \,|)}$$

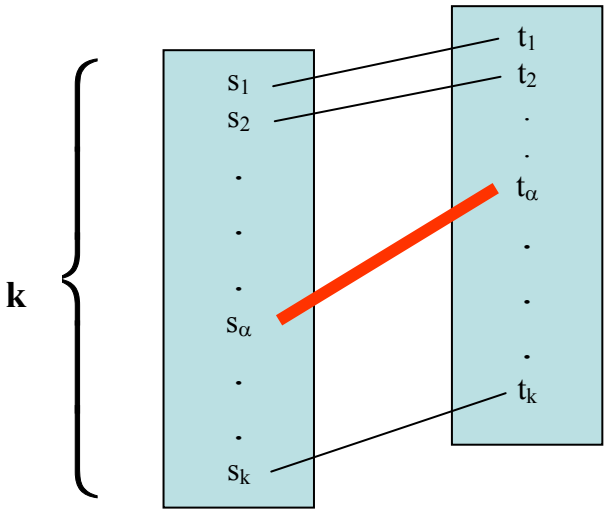

Figure 4. The "week" locality window.

If the new link starts from or ends in a token already linked, the index difference that would be null in the formula above is set to 1. This way, such candidate links

would be given support by the *LOC* feature. In the case of *chunk-based locality* the window span is given by the indexes of the first and last tokens of the chunk.

Collocation

Bi-gram lists (only content words) were built from each monolingual part of the training corpus, using the log-likelihood score (threshold of 10) and minimal occurrence frequency (3) for candidates filtering. Whenever such bigrams are found in either parts of the bitext to be aligned the respective words are dependency-linked. Then, the value of the collocation feature is computed similar to the dependency-based locality feature. The algorithm searches for the links of the lexical dependencies around the candidate link. Monolingual collocation feature is an important clue for creating multi-word alignments (one-to many, many-to-one and many-to-many).

4. COMBINING THE REIFIED ALIGNMENTS

The combination method we used in (Tufiş *et al*., 2005) was a very simple one: the alignment links produced by the base aligners were merged and the duplicates eliminated. A few heuristics (e.g. *bounded relative positions*, see (Tufiş *et al*., 2006)) further eliminated improbable links.

As mentioned before, COWAL has been rewritten to take advantage of the more principled classification techniques (such as SVM) and the methodology briefly described (Tufiş *et al*., 2006).

From a given alignment one can compute a series of properties for each of its links (such as the parameters used by the MEBA aligner). A link becomes this way a structured object that can be manipulated in various ways, independent of the bitext (or even of the lexical tokens of the link) from which it was extracted. We call this procedure *alignment reification*. The properties of the links of two or more alignments are used for our method of alignments combination.

4.1. The link classifier

We used an "off-the-shelf" solution for SVM training and classification - LIBSVM[4] (Fan *et al*., 2005) with the default parameters (C-SVC classification (soft margin) and radial basis kernel function $K(x, y) = \exp(-\gamma \|x - y\|^2)$.

We trained the classifier with both positive and negative examples of links. The links in the Gold Standard alignment (approx. 7000) were used as positive examples set. The negative examples were extracted from the alignments produced

[4] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

by COWAL, YAWA and MEBA where they differed from the Gold Standard2003. To obtain a similar number of negative examples as in the positive examples set, we automatically generated additional wrong links, by randomly changing one of the starting or the ending points of the correct links. The model of the SVM classifier, obtained by a 10-fold validation procedure, uses only the features computed by both basic aligners: *TE(S,T)*, *TE(T,S)*, *OBL(S,T)*, *LOC(S,T)*, *PA(S,T)*, *PA(T,S)*.

The alignments produced by YAWA and MEBA for the ACL 2005 test data were merged and given as input to the trained SVM classifier. The evaluation has been performed against a modified version of the ACL 2005 Gold Standard (see below).

4.2. Evaluation of the combined alignments

As mentioned before, at the shared task on English-Romanian word alignment organized at ACL 2005 in Ann Arbor, COWAL scored as the best aligner. However, after the official Gold Standard 2005 (henceforth referred to as GS1) was publicly released we noticed several alignment errors. We corrected only the indisputable errors thus obtaining a new Gold Standard, referred to as GS2. One step further was to re-tokenize the bitext (in both languages) in order to systematically deal with some compounds, various particles (this was relevant especially for Romanian: enclitic articles, double negations, clitics and clitics doubling, etc.). The initial tokenization of the GS considered as an alignment token anything delimited by white spaces. By redoing the tokenization and correspondingly updating the links, we got a second version of the Gold Standard, GS3. Table 3 displays the evaluation results, using the official evaluation tool, but the Gold Standard, GS3, for both the base and combined aligners. As compared to the results reported in (Tufiş *et al*., 2005), the *AER* figure for COWAL improved with more than 6%.

Table 3. Evaluation of the basic and the combined aligners.

Aligner	Precision	Recall	F-measure	AER
YAWA(GS3)	88.80%	74.83%	81.22%	18.78%
MEBA(GS3)	92.15%	73.40%	81.71%	18.29%
COWAL(GS3)	87.26%	80.94%	83.98%	16.02%

5. RELATED WORK

The feature-based word alignment, which we called in (Tufiş *et al*., 2006) reification, also known as *discriminative* paradigm, emerged mainly after 2005 when "several independent efforts […] demonstrated that discriminatively trained models can equal or surpass the alignment accuracy of the standard models"

(Moore *et al*., 2006). By standard models Moore, as many others, refers to word alignments obtained by the revolutionary five IBM generative probabilistic models, augmented with the so-called Model 6 (Och & Ney, 2003).

The research reported in (Lacoste-Julien *et al*., 2006) uses as features to describe an alignment link between two words: the word associations, their orthographic similarity, there relative proximity etc. The authors parameterized all the scoring functions as weighted linear combinations of feature sets. The parameter estimation was addressed by turning the estimation problem into a quadratic problem (QAP), the NP-hardness of which was overcome by a linear relaxation QAP (see (Taskar, 2004) for details). Evaluating their approach on the French-English Hansard test data from the NAACL word alignment shared task, they obtained an extremely low *AER* score of 3.8%.

In (Moore *et al*., 2006) the authors describe a two stage discriminative aligner using various features of the words in a linked pair: the words association score (a *LL* ratio score), the degree of non-monotonicity of the alignments (similar to our *OBL* feature), the number of backward jumps, association score rank, exact match (similar to *COGN*=1 feature) etc. They also use weighted linear combinations of feature sets. For training, they report on experiments using a perceptron model and an SVM model, with the latter obtaining an *AER* score of 4.7% on aligning the French-English Hansard test data from the NAACL word alignment shared task.

Although not using reification, the method described by (Liang *et al*., 2006) brings evidence that what they called *alignment by agreement*, where they optimize the agreement between the bidirectional alignments not only in the prediction stage but also during the training phase, gets better results than the standard practice of intersecting predictions of independently trained unidirectional models. On the same test data as the previous mentioned works they obtained an *AER* score of 5.2%.

Several researchers raised the issue on relevance of the *AER* score with respect to the quality of alignment and its impact on the quality of translation (Fraser & Marcu, 2007). Indeed the way *AER* was defined (Och and Ney, 2003) makes very hard to compare alignments against Gold Standards that used only S(ure) links *versus* Gold standards that used both *S*(ure) and *P*(ossible) links. This is because an alignment evaluated against a Gold Standard that has both *S* and *P* links would always obtain an *AER* better than the same alignment evaluated against a Gold Standard using only *S* links. This was the case of the English-French Gold Standard (*S* and *P* links) and the English-Romanian Gold Standard (only *S* links) used in the Word Alignment competition at ACL 2005. The best results for the two language pairs were significantly distant. Another issue that one has to consider when comparing the AER scores for two different aligners is the register and the size of the training data. HANSARD corpus is one of the largest parallel (English – French) training data and the language is somehow formulaic (parliamentary debates). At the NAACL 2003 Word Alignment competition the English – French training data, a subset of the HANSARD corpus, contained 500,000 parallel

sentences. The training corpus used for the English – Romanian shared track in the same competition contained only 45,241 parallel sentences while the language was much freer (journalistic register). In the previously mentioned paper, Fraser and Marcu claim that the impact of *AER* score improvement is unclear and they argue in favour of a new metric, based on a correlation coefficient (squared Pearson product-moment coefficient). However, in (Haghighi *et al*., 2009, Table 3) the authors bring clear evidence of correlation between improvement of the alignment and the increased quality of the translation.

6. CONCLUSIONS AND FURTHER WORK

Neither YAWA nor MEBA needs an a priori bilingual dictionary, as this will be automatically extracted by either TREQ-AL or GIZA++. We made evaluation of the individual alignments in both experimental settings: without a start-up bilingual lexicon and with an initial mid-sized bilingual lexicon. Surprisingly enough, we found that while the performance of YAWA increases a little bit (approx. 1% increase of the F-measure), MEBA is doing better without an additional lexicon. Therefore, in the evaluation presented in section 4, MEBA uses only the training data vocabulary.

YAWA is very sensitive to the quality of the bilingual lexicons it uses. We used automatically extracted translation lexicons (with or without a seed lexicon), and the noise inherently present might have had a bad influence on YAWA's precision. Replacing the automated generated bilingual lexicons with validated (reference) bilingual lexicons) would further improve the overall performance of this aligner. Yet, this might be a harder to meet condition for some pairs of languages than using parallel corpora. A new version of YAWA, incorporating dependency linking information, is almost finished and its individual or embedded into COWAL evaluation will be reported in a future paper.

MEBA is more versatile and it is not as sensitive as YAWA to the quality of the translation lexicons but, on the other hand, it is very sensitive to the values of the parameters that control its behaviour. Currently they are set according to the developers' intuition and after the analysis of the results from several trials. Since this activity is pretty time consuming, we plan to extend MEBA with a supervised learning module, which would automatically determine the "optimal" parameters (thresholds and weights) values.

Acknowledgements. The work reported here is currently funded by the STAR project, financed by the Romanian National Council for Scientific Research (CNCSIS) under the grant no. ID_1443.

REFERENCES

1. AHRENBERG L., ANDERSSON M., and MERKEL M. A., Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts, in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pp. 29-35. Université de Montréal. 1998.
2. AL-ONAIZAN Y., CURIN J., JAHR M., KNIGHT K., LAFFERTY J., MELAMED D., OCH F.J, PURDY D., SMITH N.A., and YAROWSKY D., Statistical Machine Translation: Final Report. *Johns Hopkins University 1999 Summer Workshop on Language Engineering*, Center for Language and Speech Processing, Baltimore, MD, USA. 1999.
3. BREW C. and MCKELVIE D., Word-pair extraction for lexicography, in Kemal Oflazer and Harold Somers (eds.) *Proceedings of the Second International Conference on New Methods in Language Processing*, pages 45–55, Ankara, September. Bilkent University. 1996.
4. BROWN P., F., DELLA PIETRA S., A., DELLA PIETRA V.J., and MERCER R.J., The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2): 263–311. 1993.
5. CEAUŞU A., ŞTEFĂNESCU D., and TUFIŞ D., Acquis Communautaire sentence alignment using Support Vector Machines, in *Proceedings of the 5th LREC Conference*, Genoa, Italy, 22–28 May. 2006.
6. CEAUŞU A., Maximum Entropy Tiered Tagging, in Janneke Huitink & Sophia Katrenko (eds.), *Proceedings of the Eleventh ESSLLI Student Session, ESSLLI 2006*, pp. 173–179. 2006.
7. DIETTERICH, T.G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation,* 10(7):1895–1924. 1998.
8. DIMITROVA L., ERJAVEC T., IDE N., KAALEP H.J., PETKEVIC V., and TUFIŞ D., Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In Christian Boitet and Pete Whitelock (eds.), *Proceedings of the Joint 17th International Conference on Computational Linguistics 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 1998)*, pp. 315–319, Montreal, Canada, Morgan Kaufmann Publishers. 1998.
9. DUNNING T., Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1):61-74. 1993.
10. FAN R., CHEN P.-H., and LIN C.J., *Working set selection using the second order information for training SVM*. Technical report, Department of Computer Science, National Taiwan University. 2005. (www.csie.ntu.edu.tw/~cjlin/papers/ quadworkset.pdf).
11. FRASER A. and MARCU D., Measuring Word Alignment Quality for Statistical Machine Translation. *Computational Linguistics*, 33(3): 293–303. 2007.
12. GALE W.A. and CHURCH K.W., Identifying word correspondences in parallel texts, in *Proceedings of the Fourth DARPA Workshop on Speech and Natural Language*. Asilomar, CA, pp. 152–157. 1991.
13. GILDEA D., Loosely Tree-Based Alignment for Machine Translation, in *Proceedings of the 41st ACL*, Sapporo, Japan, pp. 80–87. 2003.
14. HIEMSTRA D., Deriving a bilingual lexicon for cross language information retrieval, in *Proceedings of Gronics*, pp. 21–26. 1997
15. ITTYCHERIAH A. and ROUKOS S., A maximum entropy word aligner for Arabic-English machine translation, in *Proceedings of HLT-EMNLP*, Vancouver, Canada, pp. 89–96. 2005.
16. ION R., CEAUŞU A., and TUFIŞ D., Dependency-Based Phrase Alignment. In *Proceedings of the 5th LREC Conference*, Genoa, Italy. 2006.
17. ION R., *Word Sense Disambiguation Methods Applied to English and Romanian* (in Romanian). PhD Thesis, Research Institute for Artificial Intelligence, Romanian Academy, Bucharest, 143 pages. 2007.
18. KAY M. and RÖSCHEISEN M., Text-translation alignment. *Computational Linguistics*, 19(1), Special issue on using large corpora: I, ISSN:0891-2017, pp. 121–142, 1993.

19. KOEHN, P., OCH, F.J., and MARCU, D. Statistical Phrase-Based Translation, in Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference NAACL/HLT 2003, Edmonton, Canada. 2003.
20. KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A., and HERBST E., Moses: Open Source Toolkit for Statistical Machine Translation, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2007)*, demonstration session, Prague, Czech Republic, June 2007.
21. KUPIEC J., An algorithm for finding noun phrase correspondences in bilingual corpora, in *Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics*, pp. 17–22. 1993.
22. LACOSTE-JULIEN S., TASKAR B., KLEIN D., and JORDAN M.L., Word Alignment via Quadratic Assignment. In *Proceedings of the NAACL HLT Conference*, New York City, pp. 112–119. 2006.
23. LIANG P., TASKAR B., and KLEIN D., Alignment by agreement, in *Proceedings of the NAACL HLT Conference*, New York City, pp. 104–111. 2006.
24. MARTIN M., MIHALCEA R., and PEDERSEN T., Word Alignment for Languages with Scarce Resources, in *Proceedings of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond"*, Ann Arbor, Michigan, Association for Computational Linguistics, pp. 65–74. 2005.
25. MELAMED D., *Empirical Methods for Exploiting Parallel Texts*. Cambridge, MA, MIT Press. 2001.
26. MEYERS A., YANGARBER R., and GRISHAM R., Alignment of shared forests for bilingual corpora, in *Proceedings of 16th International Conference on Computational Linguistics COLING-96*, Copenhagen, Denmark, pp. 460–465. 1996
27. MIHALCEA R. and PEDERSEN T., An Evaluation Exercise for Word Alignment. In *Proceedings of the HLT-NAACL 2003 Workshop: Building and Using Parallel Texts Data Driven Machine Translation and Beyond*. Edmonton, Canada, pp. 1–10. 2003.
28. MOORE R.C., Fast and Accurate Sentence Alignment of Bilingual Corpora in Machine Translation: From Research to Real Users, in *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas*, Tiburon, California, Springer-Verlag, Heidelberg, Germany, pp. 135–244. 2002.
29. MOORE R.C., YIH W., and BODE A., Improved Discriminative Bilingual Word Alignment, in *Proceedings of the 21st COLING and the 44th Annual Meeting of ACL*, Sydney, pp. 513–520. 2006.
30. OCH F.J. and NEY H., A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51. 2003.
31. OCH F.J. and NEY H., Improved Statistical Alignment Models, in *Proceedings of the 38th Conference of ACL*, Hong Kong, pp. 440–447. 2000
32. SMADJA F., MCKEOWN K.R., and HATZIVASSILOGLOU V., Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1), 1:38. 1996
33. TASKAR B., Learning Structured Prediction Models: A Large Margin Approach. PhD. Thesis, Stanford University. 2004.
34. TIEDEMANN J., Combining clues for word alignment, in *Proceedings of the 10th EACL*, Budapest, Hungary, pp. 339–346. 2003.
35. TUFIŞ D., Tiered Tagging and Combined Language Models Classifiers, in Václav Matousek, Pavel Mautner, Jana Ocelíková, and Petr Sojka (eds.), *Text, Speech and Dialogue (TSD 1999)*, Lecture Notes in Artificial Intelligence 1692, pp. 28–33. Springer Berlin / Heidelberg. ISBN 978-3-540-66494-9. 1999
36. TUFIŞ D., A cheap and fast way to build useful translation lexicons. In *Proceedings of COLING 2002*, Taipei, China, pp. 1030–1036. 2002.
37. TUFIŞ D., ION R., CEAUŞU A., and ŞTEFĂNESCU D., Combined Aligners. In *Proceeding of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond*", Ann Arbor, Michigan, Association for Computational Linguistics, pp. 107–110. 2005.

38. TUFIŞ D., ION R., CEAUŞU A., and ŞTEFĂNESCU D., Improved Lexical Alignment by Combining Multiple Reified Alignments, in *Proceedings of the Conference of European Association for Computational Linguistics, EACL2006*, Trento, 3–7 April, pp. 145–152. 2006.
39. TUFIŞ D., ION R., CEAUŞU A., and ŞTEFĂNESCU D., RACAI's Linguistic Web Services, in *Proceedings of the 6th Language Resources and Evaluation Conference – LREC2008*, Marrakech, Morocco, May. ELRA – European Language Resources Association. ISBN 2-9517408-4-0. 2008.
40. YAMADA K. and KNIGHT K., A syntax-based statistical translation model, in Proceedings *of the 39th Meeting of the Association for Computational Linguistics ACL 2001*, Toulouse, France, pp. 523–530. 2001.
41. ZIPF G.K., *The Psycho-biology of Language: An Introduction to Dynamic Philology*. Routledge, London, UK. 1936.