An Evaluation Exercise for Word Alignment

Rada Mihalcea
Department of Computer Science
University of North Texas
Denton, TX 76203
rada@cs.unt.edu

Ted Pedersen
Department of Computer Science
University of Minnesota
Duluth, MN 55812
tpederse@umn.edu

Abstract

This paper presents the task definition, resources, participating systems, and comparative results for the shared task on word alignment, which was organized as part of the HLT/NAACL 2003 Workshop on Building and Using Parallel Texts. The shared task included Romanian-English and English-French sub-tasks, and drew the participation of seven teams from around the world.

1 Defining a Word Alignment Shared Task

The task of word alignment consists of finding correspondences between words and phrases in parallel texts. Assuming a sentence aligned bilingual corpus in languages L1 and L2, the task of a word alignment system is to indicate which word token in the corpus of language L1 corresponds to which word token in the corpus of language L2.

As part of the HLT/NAACL 2003 workshop on "Building and Using Parallel Texts: Data Driven Machine Translation and Beyond", we organized a shared task on word alignment, where participating teams were provided with training and test data, consisting of sentence aligned parallel texts, and were asked to provide automatically derived word alignments for all the words in the test set. Data for two language pairs were provided: (1) English-French, representing languages with rich resources (20 million word parallel texts), and (2) Romanian-English, representing languages with scarce resources (1 million word parallel texts). Similar with the Machine Translation evaluation exercise organized by NIST[1], two subtasks were defined, with teams being encouraged to participate in both subtasks.

[1] http://www.nist.gov/speech/tests/mt/

1. *Limited resources*, where systems are allowed to use **only** the resources provided.

2. *Unlimited resources*, where systems are allowed to use any resources in addition to those provided. Such resources had to be explicitly mentioned in the system description.

Test data were released one week prior to the deadline for result submissions. Participating teams were asked to produce word alignments, following a common format as specified below, and submit their output by a certain deadline. Results were returned to each team within three days of submission.

1.1 Word Alignment Output Format

The word alignment result files had to include one line for each word-to-word alignment. Additionally, lines in the result files had to follow the format specified in Fig.1.

While the $S|P$ and confidence fields overlap in their meaning, the intent of having both fields available is to enable participating teams to draw their own line on what they consider to be a Sure or Probable alignment. Both these fields were optional, with some standard values assigned by default.

1.1.1 A Running Word Alignment Example

Consider the following two aligned sentences:
[English] *<s snum=18> They had gone . </s>*
[French] *<s snum=18> Ils etaient alles . </s>*
A correct word alignment for this sentence is
18 1 1
18 2 2
18 3 3
18 4 4
stating that: all the word alignments pertain to sentence 18, the English token 1 *They* aligns with the French token 1 *Ils*, the English token 2 *had*, aligns with the French token 2 *etaient*, and so on. Note that punctuation is also

sentence_no position_L1 position_L2 [$S|P$] [confidence]

where:

- **sentence_no** represents the id of the sentence within the test file. Sentences in the test data already have an id assigned. (see the examples below)
- **position_L1** represents the position of the token that is aligned from the text in language L1; the first token in each sentence is token 1. (not 0)
- **position_L2** represents the position of the token that is aligned from the text in language L2; again, the first token is token 1.
- $S|P$ can be either S or P, representing a Sure or Probable alignment. All alignments that are tagged as S are also considered to be part of the P alignments set (that is, all alignments that are considered "Sure" alignments are also part of the "Probable" alignments set). If the $S|P$ field is missing, a value of S will be assumed by default.
- **confidence** is a real number, in the range (0-1] (1 meaning highly confident, 0 meaning not confident); this field is optional, and by default confidence number of 1 was assumed.

Figure 1: Word Alignment file format

aligned (English token 4 aligned with French token 4), and counts towards the final evaluation figures.

Alternatively, systems could also provide an $S|P$ marker and/or a confidence score, as shown in the following example:

18 1 1 1
18 2 2 P 0.7
18 3 3 S
18 4 4 S 1

with missing $S|P$ fields considered by default to be S, and missing confidence scores considered by default 1.

1.2 Annotation Guide for Word Alignments

The annotation guide and illustrative word alignment examples were mostly drawn from the Blinker Annotation Project. Please refer to (Melamed, 1999, pp. 169–182) for additional details.

1. All items separated by a white space are considered to be a word (or token), and therefore have to be aligned. (punctuation included)

2. Omissions in translation use the NULL token, i.e. token with id 0. For instance, in the examples below:

   [English]: *<s snum=18> And he said , appoint me thy wages , and I will give it . </s>*

   [French]: *<s snum=18> fixe moi ton salaire , et je te le donnerai . </s>*

   and he said from the English sentence has no corresponding translation in French, and therefore all these words are aligned with the token id 0.

   ...
   18 1 0
   18 2 0
   18 3 0
   18 4 0
   ...

3. Phrasal correspondences produce multiple word-to-word alignments. For instance, in the examples below:

   English: *<s snum=18> cultiver la terre </s>*

   French: *<s snum=18> to be a husbandman </s>*

   since the words do not correspond one to one, and yet the two phrases mean the same thing in the given context, the phrases should be linked as wholes, by linking each word in one to each word in another. For the example above, this translates into 12 word-to-word alignments:

18 1 1	18 1 2
18 1 3	18 1 4
18 2 1	18 2 2
18 2 3	18 2 4
18 3 1	18 3 2
18 3 3	18 3 4

2 Resources

The shared task included two different language pairs: the alignment of words in English-French parallel texts, and in Romanian-English parallel texts. For each language pair, training data were provided to participants. Systems relying only on these resources were considered part of the *Limited Resources* subtask. Systems making use of any additional resources (e.g. bilingual dictionaries, additional parallel corpora, and others) were classified under the *Unlimited Resources* category.

2.1 Training Data

Two sets of training data were made available.

1. A set of Romanian-English parallel texts, consisting of about 1 million Romanian words, and about the same number of English words. These data consisted of:

   - Parallel texts collected from the Web using a semi-supervised approach. The URLs format for pages containing potential parallel translations were manually identified (mainly from the archives of Romanian newspapers). Next, texts were automatically downloaded and sentence aligned. A manual verification of the

alignment was also performed. These data collection process resulted in a corpus of about 850,000 Romanian words, and about 900,000 English words.

- Orwell's 1984, aligned within the MULTEXT-EAST project (Erjavec et al., 1997), with about 130,000 Romanian words, and a similar number of English words.
- The Romanian Constitution, for about 13,000 Romanian words and 13,000 English words.

2. A set of English-French parallel texts, consisting of about 20 million English words, and about the same number of French words. This is a subset of the Canadian Hansards, processed and sentence aligned by Ulrich Germann at ISI (Germann, 2001).

All data were pre-tokenized. For English and French, we used a version of the tokenizers provided within the EGYPT Toolkit[2]. For Romanian, we used our own tokenizer. Identical tokenization procedures were used for training, trial, and test data.

2.2 Trial Data

Two sets of trial data were made available at the same time training data became available. Trial sets consisted of sentence aligned texts, provided together with manually determined word alignments. The main purpose of these data was to enable participants to better understand the format required for the word alignment result files. Trial sets consisted of 37 English-French, and 17 Romanian-English aligned sentences.

2.3 Test Data

A total of 447 English-French aligned sentences (Och and Ney, 2000), and 248 Romanian-English aligned sentences were released one week prior to the deadline. Participants were required to run their word alignment systems on these two sets, and submit word alignments. Teams were allowed to submit an unlimited number of results sets for each language pair.

2.3.1 Gold Standard Word Aligned Data

The gold standard for the two language pair alignments were produced using slightly different alignment procedures, which allowed us to study different schemes for producing gold standards for word aligned data.

For English-French, annotators where instructed to assign a Sure or Probable tag to each word alignment they produced. The intersection of the Sure alignments produced by the two annotators led to the final Sure aligned set, while the reunion of the Probable alignments led to the final Probable aligned set. The Sure alignment set is guaranteed to be a subset of the Probable alignment set. The annotators did not produce any NULL alignments. Instead, we assigned NULL alignments as a default back-up mechanism, which forced each word to belong to at least one alignment. The English-French aligned data were produced by Franz Och and Hermann Ney (Och and Ney, 2000).

For Romanian-English, annotators were instructed to assign an alignment to *all* words, with specific instructions as to when to assign a NULL alignment. Annotators were not asked to assign a Sure or Probable label. Instead, we had an arbitration phase, where a third annotator judged the cases where the first two annotators disagreed. Since an inter-annotator agreement was reached for all word alignments, the final resulting alignments were considered to be Sure alignments.

3 Evaluation Measures

Evaluations were performed with respect to four different measures. Three of them – precision, recall, and F-measure – represent traditional measures in Information Retrieval, and were also frequently used in previous word alignment literature. The fourth measure was originally introduced by (Och and Ney, 2000), and proposes the notion of *quality of word alignment*.

Given an alignment $\mathcal{A}$, and a gold standard alignment $\mathcal{G}$, each such alignment set eventually consisting of two sets $\mathcal{A}_\mathcal{S}$, $\mathcal{A}_\mathcal{P}$, and $\mathcal{G}_\mathcal{S}$, $\mathcal{G}_\mathcal{P}$ corresponding to Sure and Probable alignments, the following measures are defined (where T is the alignment type, and can be set to either S or P).

$$P_T = \frac{|A_T \cap G_T|}{|A_T|} \tag{1}$$

$$R_T = \frac{|A_T \cap G_T|}{|G_T|} \tag{2}$$

$$F_T = \frac{2P_T R_T}{P_T + R_T} \tag{3}$$

$$AER = 1 - \frac{|A_P \cap G_S| + |A_P \cap G_P|}{|A_P| + |G_S|} \tag{4}$$

Each word alignment submission was evaluated in terms of the above measures. Moreover, we conducted two sets of evaluations for each submission:

- NULL-Align, where each word was enforced to belong to at least one alignment; if a word did not belong to any alignment, a NULL Probable alignment was assigned by default. This set of evaluations pertains to *full coverage* word alignments.
- NO-NULL-Align, where all NULL alignments were removed from both submission file and gold standard data.

[2] http://www.clsp.jhu.edu/ws99/projects/mt/toolkit/

Team	System name	Description
Language Technologies Institute, CMU	BiBr	(Zhao and Vogel, 2003)
MITRE Corporation	Fourday	(Henderson, 2003)
RALI - Université the Montréal	Ralign	(Simard and Langlais, 2003)
Romanian Academy Institute of Artificial Intelligence	RACAI	(Tufiş et al., 2003)
University of Alberta	ProAlign	(Lin and Cherry, 2003)
University of Minnesota, Duluth	UMD	(Thomson McInnes and Pedersen, 2003)
Xerox Research Centre Europe	XRCE	(Dejean et al., 2003)

Table 1: Teams participating in the word alignment shared task

We conducted therefore 14 evaluations for each submission file: AER, Sure/Probable Precision, Sure/Probable Recall, and Sure/Probable F-measure, with a different figure determined for NULL-Align and NO-NULL-Align alignments.

4 Participating Systems

Seven teams from around the world participated in the word alignment shared task. Table 1 lists the names of the participating systems, the corresponding institutions, and references to papers in this volume that provide detailed descriptions of the systems and additional analysis of their results.

All seven teams participated in the Romanian-English subtask, and five teams participated in the English-French subtask.[3] There were no restrictions placed on the number of submissions each team could make. This resulted in a total of 27 submissions from the seven teams, where 14 sets of results were submitted for the English-French subtask, and 13 for the Romanian-English subtask. Of the 27 total submissions, there were 17 in the *Limited resources* subtask, and 10 in the *Unlimited resources* subtask. Tables 2 and 3 show all of the submissions for each team in the two subtasks, and provide a brief description of their approaches.

While each participating system was unique, there were a few unifying themes.

Four teams had approaches that relied (to varying degrees) on an IBM model of statistical machine translation (**?**). UMD was a straightforward implementation of IBM Model 2, BiBr employed a boosting procedure in deriving an IBM Model 1 lexicon, Ralign used IBM Model 2 as a foundation for their recursive splitting procedure, and XRCE used IBM Model 4 as a base for alignment with lemmatized text and bilingual lexicons.

Two teams made use of syntactic structure in the text to be aligned. ProAlign satisfies constraints derived from a dependency tree parse of the English sentence being aligned. BiBr also employs syntactic constraints that must be satisfied. However, these come from parallel text that has been shallowly parsed via a method known as bilingual bracketing.

Three teams approached the shared task with baseline or prototype systems. Fourday combines several intuitive baselines via a nearest neighbor classifier, RACAI carries out a greedy alignment based on an automatically extracted dictionary of translations, and UMD's implementation of IBM Model 2 provides an experimental platform for their future work incorporating prior knowledge about cognates. All three of these systems were developed within a short period of time before and during the shared task.

5 Results and Discussion

Tables 4 and 5 list the results obtained by participating systems in the Romanian-English task. Similarly, results obtained during the English-French task are listed in Tables 6 and 7.

For Romanian-English, limited resources, XRCE systems (XRCE.Nolem-56k.RE.2 and XRCE.Trilex.RE.3) seem to lead to the best results. These are systems that are based on GIZA++, with or without additional resources (lemmatizers and lexicons). For unlimited resources, ProAlign.RE.1 has the best performance.

For English-French, Ralign.EF.1 has the best performance for limited resources, while ProAlign.EF.1 has again the largest number of top ranked figures for unlimited resources.

To make a cross-language comparison, we paid particular attention to the evaluation of the Sure alignments, since these were collected in a similar fashion (an agreement had to be achieved between two different annotators). The results obtained for the English-French Sure alignments are significantly higher (80.54% best F-measure) than those for Romanian-English Sure alignments (71.14% best F-measure). Similarly, AER for English-French (5.71% highest error reduction) is clearly better than the AER for Romanian-English (28.86% highest error reduction).

This difference in performance between the two data sets is not a surprise. As expected, word alignment, like

[3]The two teams that did not participate in English-French were Fourday and RACAI.

many other NLP tasks (Banko and Brill, 2001), highly benefits from large amounts of training data. Increased performance is therefore expected when larger training data sets are available.

The only evaluation set where Romanian-English data leads to better performance is the Probable alignments set. We believe however that these figures are not directly comparable, since the English-French Probable alignments were obtained as a reunion of the alignments assigned by two different annotators, while for the Romanian-English Probable set two annotators had to reach an agreement (that is, an intersection of their individual alignment assignments).

Interestingly, in an overall evaluation, the limited resources systems seem to lead to better results than those with unlimited resources. Out of 28 different evaluation figures, 20 top ranked figures are provided by systems with limited resources. This suggests that perhaps using a large number of additional resources does not seem to improve a lot over the case when only parallel texts are employed.

Ranked results for all systems are plotted in Figures 2 and 3. In the graphs, systems are ordered based on their AER scores. System names are preceded by a marker to indicate the system type: L stands for *Limited Resources*, and U stands for *Unlimited Resources*.

6 Conclusion

A shared task on word alignment was organized as part of the HLT/NAACL 2003 Workshop on Building and Using Parallel Texts. In this paper, we presented the task definition, and resources involved, and shortly described the participating systems. The shared task included Romanian-English and English-French sub-tasks, and drew the participation of seven teams from around the world. Comparative evaluations of results led to interesting insights regarding the impact on performance of (1) various alignment algorithms, (2) large or small amounts of training data, and (3) type of resources available. Data and evaluation software used in this exercise are available online at http://www.cs.unt.edu/~rada/wpt.

Acknowledgments

There are many people who contributed greatly to making this word alignment evaluation task possible. We are grateful to all the participants in the shared task, for their hard work and involvement in this evaluation exercise. Without them, all these comparative analyses of word alignment techniques would not be possible.

We are very thankful to Franz Och from ISI and Hermann Ney from RWTH Aachen for kindly making their English-French word aligned data available to the workshop participants; the Hansards made available by Ulrich Germann from ISI constituted invaluable data for the English-French shared task. We would also like to thank the student volunteers from the Department of English, Babes-Bolyai University, Cluj-Napoca, Romania who helped creating the Romanian-English word aligned data.

We are also grateful to all the Program Committee members of the current workshop, for their comments and suggestions, which helped us improve the definition of this shared task. In particular, we would like to thank Dan Melamed for suggesting the two different subtasks (limited and unlimited resources), and Michael Carl and Phil Resnik for initiating interesting discussions regarding phrase-based evaluations.

References

M. Banko and E. Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Lingusitics (ACL-2001)*, Toulouse, France, July.

Herve Dejean, Eric Gaussier, Cyril Goutte, and Kenji Yamada. 2003. Reducing parameter space for word alignment. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 23–26, Edmonton, Alberta, Canada, May 31. Association for Computational Linguistics.

T. Erjavec, N. Ide, and D. Tufis. 1997. Encoding and parallel alignment of linguistic corpora in six central and Eastern European languages. In *Proceedings of the Joint ACH/ALL Conference*, Queen's University, Kingston, Ontario, June.

U. Germann. 2001. Aligned hansards of the 36th parliament of canada. http://www.isi.edu/natural-language/download/hansard/.

John C. Henderson. 2003. Word alignment baselines. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 27–30, Edmonton, Alberta, Canada, May 31. Association for Computational Linguistics.

Dekang Lin and Colin Cherry. 2003. Proalign: Shared task system description. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 11–14, Edmonton, Alberta, Canada, May 31. Association for Computational Linguistics.

D.I. Melamed. 1999. *Empirical Methods for Exploiting Parallel Texts*. MIT Press.

F. Och and H. Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-ACL 2000)*, Saarbrucken, Germany, August.

Michel Simard and Philippe Langlais. 2003. Statistical translation alignment with compositionality constraints. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 19–22, Edmonton, Alberta, Canada, May 31. Association for Computational Linguistics.

Bridget Thomson McInnes and Ted Pedersen. 2003. The duluth word alignment system. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 40–43, Edmonton, Alberta, Canada, May 31. Association for Computational Linguistics.

Dan Tufiş, Ana-Maria Barbu, and Radu Ion. 2003. Treqal: A word alignment system with limited language resources. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 36–39, Edmonton, Alberta, Canada, May 31. Association for Computational Linguistics.

Bing Zhao and Stephan Vogel. 2003. Word alignment based on bilingual bracketing. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 15–18, Edmonton, Alberta, Canada, May 31. Association for Computational Linguistics.

System	Resources	Description
BiBr.EF.1	Limited	baseline of bilingual bracketing
BiBr.EF.2	Unlimited	baseline of bilingual bracketing + English POS tagging
BiBr.EF.3	Unlimited	baseline of bilingual bracketing + English POS tagging and base NP
BiBr.EF.4	Limited	reverse direction of BiBr.EF.1
BiBr.EF.5	Unlimited	reverse direction of BiBr.EF.2
BiBr.EF.6	Unlimited	reverse direction of BiBr.EF.3
BiBr.EF.7	Limited	intersection of BiBr.EF.1 & BiBr.EF.3
BiBr.EF.8	Unlimited	intersection of BiBr.EF.3 & BiBr.EF.6
ProAlign.EF.1	Unlimited	cohesion between source and target language + English parser + distributional similarity for English words
Ralign.EF.1	Limited	Giza (IBM Model 2) + recursive parallel segmentation
UMD.EF.1	Limited	IBM Model 2, trained with 1/20 of the corpus, distortion 2, iterations 4
XRCE.Base.EF.1	Limited	GIZA++ (IBM Model 4) with English and French lemmatizer
XRCE.Nolem.EF.2	Limited	GIZA++ only (IBM Model 4), trained with 1/4 of the corpus
XRCE.Nolem.EF.3	Limited	GIZA++ only (IBM Model 4), trained with 1/2 of the corpus

Table 2: Short description for English-French systems

System	Resources	Description
BiBr.RE.1	Limited	baseline of bilingual bracketing
BiBr.RE.2	Unlimited	baseline of bilingual bracketing + English POS tagging
BiBr.RE.3	Unlimited	baseline of bilingual bracketing + English POS tagging and base NP
Fourday.RE.1	Limited	nearest neighbor combination of baseline measures
ProAlign.RE.1	Unlimited	cohesion between source and target language + English parser + distributional similarity for English words
RACAI.RE.1	Unlimited	translation equivalence dictionary (**?**) + POS tagging
Ralign.RE.1	Limited	Giza (IBM Model 2) + recursive parallel segmentation
UMD.RE.1	Limited	IBM Model 2, trained with all the corpus, distortion 4, iterations 4
UMD.RE.2	Limited	IBM Model 2, trained with all the corpus, distortion 2, iterations 4
XRCE.Base.RE.1	Limited	GIZA++ (IBM Model 4), with English lemmatizer
XRCE.Nolem.RE.2	Limited	GIZA++ only (IBM Model 4)
XRCE.Trilex.RE.3	Limited	GIZA++ only (IBM Model 4), with English lemmatizer and trinity lexicon
XRCE.Trilex.RE.4	Limited	GIZA++ only (IBM Model 4), with English lemmatizer and trinity lexicon

Table 3: Short description for Romanian-English systems

System	P_S	R_S	F_S	P_P	R_P	F_P	AER
Limited Resources							
BiBr.RE.1	70.65%	55.75%	62.32%	59.60%	57.65%	58.61%	41.39%
Fourday.RE.1	0.00%	0.00%	0.00%	52.83%	42.86%	47.33%	52.67%
Ralign.RE.1	**92.00%**	45.06%	60.49%	63.63%	**65.92%**	64.76%	35.24%
UMD.RE.1	57.67%	49.70%	53.39%	57.67%	49.70%	53.39%	46.61%
UMD.RE.2	58.29%	49.99%	53.82%	58.29%	49.99%	53.82%	46.18%
XRCE.Base.RE.1	79.28%	61.14%	69.03%	79.28%	61.14%	**69.03%**	30.97%
XRCE.Nolem-56K.RE.2	82.65%	**62.44%**	**71.14%**	**82.65%**	62.44%	71.14%	**28.86%**
XRCE.Trilex.RE.3	80.97%	61.89%	70.16%	80.97%	61.89%	70.16%	29.84%
XRCE.Trilex.RE.4	79.76%	61.31%	69.33%	79.76%	61.31%	69.33%	30.67%
Unlimited Resources							
BiBr.RE.2	70.46%	55.51%	62.10%	58.40%	57.59%	57.99%	41.39%
BiBr.RE.3	70.36%	55.47%	62.04%	58.17%	58.12%	58.14%	41.86%
RACAI.RE.1	81.29%	**60.26%**	69.21%	81.29%	**60.26%**	69.21%	30.79%
ProAlign.RE.1	**88.22%**	58.91%	**70.64%**	**88.22%**	58.91%	**70.64%**	**29.36%**

Table 4: Results for Romanian-English, NO-NULL-Align

System	P_S	R_S	F_S	P_P	R_P	F_P	AER
Limited Resources							
BiBr.RE.1	70.65%	48.32%	57.39%	57.38%	52.62%	54.90%	45.10%
Fourday.RE.1	0.00%	0.00%	0.00%	35.85%	45.88%	40.25%	59.75%
Ralign.RE.1	**92.00%**	39.05%	54.83%	63.63%	57.13%	60.21%	39.79%
UMD.RE.1	56.21%	43.17%	48.84%	45.51%	47.76%	46.60%	53.40%
UMD.RE.2	56.58%	43.45%	49.15%	46.00%	47.88%	46.92%	53.08%
XRCE.Base.RE.1	79.28%	52.98%	63.52%	61.59%	61.50%	61.54%	38.46%
XRCE.Nolem-56K.RE.2	82.65%	**54.12%**	**65.41%**	61.59%	61.50%	61.54%	38.46%
XRCE.Trilex.RE.3	80.97%	53.64%	64.53%	**63.64%**	**61.58%**	**62.59%**	**37.41%**
XRCE.Trilex.RE.4	79.76%	53.14%	63.78%	62.22%	61.37%	61.79%	38.21%
Unlimited Resources							
BiBr.RE.2	70.46%	48.11%	57.18%	56.01%	52.26%	54.07%	45.93%
BiBr.RE.3	70.36%	48.08%	57.12%	56.05%	52.87%	54.42%	45.58%
RACAI.RE.1	60.30%	**62.38%**	61.32%	59.87%	**62.42%**	61.12%	38.88%
ProAlign.RE.1	**88.22%**	51.06%	**64.68%**	**61.71%**	62.05%	**61.88%**	**38.12%**

Table 5: Results for Romanian-English, NULL-Align

System	P_S	R_S	F_S	P_P	R_P	F_P	AER
Limited Resources							
BiBr.EF.1	49.85%	79.45%	61.26%	67.23%	29.24%	40.76%	28.23%
BiBr.EF.4	51.46%	82.42%	63.36%	66.65%	32.68%	43.86%	28.01%
BiBr.EF.7	63.03%	74.59%	68.32%	66.11%	30.06%	41.33%	29.38%
Ralign.EF.1	**72.54%**	80.61%	**76.36%**	77.56%	**38.19%**	**51.18%**	18.50%
UMD.EF.1	37.98%	64.66%	47.85%	59.69%	23.53%	33.75%	38.47%
XRCE.Base.EF.1	50.89%	84.67%	63.57%	83.22%	32.05%	46.28%	16.23%
XRCE.Nolem.EF.2	55.54%	**93.46%**	69.68%	89.65%	34.92%	50.27%	8.93%
XRCE.Nolem.EF.3	55.43%	93.81%	69.68%	**90.09%**	35.30%	50.72%	**8.53%**
Unlimited Resources							
BiBr.EF.2	50.05%	79.89%	61.54%	66.92%	29.14%	40.60%	28.24%
BiBr.EF.3	50.21%	80.26%	61.80%	63.79%	30.52%	41.29%	30.38%
BiBr.EF.5	51.27%	82.17%	63.15%	67.22%	32.56%	43.87%	27.71%
BiBr.EF.6	51.91%	83.26%	63.95%	62.21%	**34.58%**	**44.45%**	31.32%
BiBr.EF.8	66.34%	74.86%	70.34%	61.62%	31.37%	41.57%	32.48%
ProAlign.EF.1	**71.94%**	**91.48%**	**80.54%**	**96.49%**	28.41%	43.89%	**5.71%**

Table 6: Results for English-French, NO-NULL-Align

System	P_S	R_S	F_S	P_P	R_P	F_P	AER
Limited Resources							
BiBr.EF.1	49.85%	79.45%	61.26%	60.32%	29.12%	39.28%	33.37%
BiBr.EF.4	51.46%	82.42%	63.36%	61.64%	32.41%	42.48%	31.91%
BiBr.EF.7	**63.03%**	74.59%	68.32%	51.35%	30.45%	38.23%	40.97%
Ralign.EF.1	72.54%	80.61%	**76.36%**	**77.56%**	**36.79%**	**49.91%**	**18.50%**
UMD.EF.1	37.19%	64.66%	47.22%	41.93%	24.08%	30.59%	51.71%
XRCE.Base.EF.1	50.89%	84.67%	63.57%	64.96%	32.73%	43.53%	28.99%
XRCE.Nolem.EF.2	55.54%	93.46%	69.68%	70.98%	35.61%	47.43%	22.10%
XRCE.Nolem.EF.3	55.43%	**93.81%**	69.68%	72.01%	36.00%	48.00%	21.27%
Unlimited Resources							
BiBr.EF.2	50.05%	79.89%	61.54%	59.89%	28.96%	39.04%	33.48%
BiBr.EF.3	50.21%	80.26%	61.80%	57.85%	30.28%	39.75%	35.03%
BiBr.EF.5	51.27%	82.17%	63.15%	**62.05%**	32.23%	42.43%	**31.69%**
BiBr.EF.6	51.91%	83.26%	63.95%	58.41%	**34.20%**	**43.14%**	34.47%
BiBr.EF.8	66.34%	74.86%	70.34%	48.50%	31.76%	38.38%	43.37%
ProAlign.EF.1	**71.94%**	**91.48%**	**80.54%**	56.02%	30.05%	39.62%	33.71%

Table 7: Results for English-French, NULL-Align

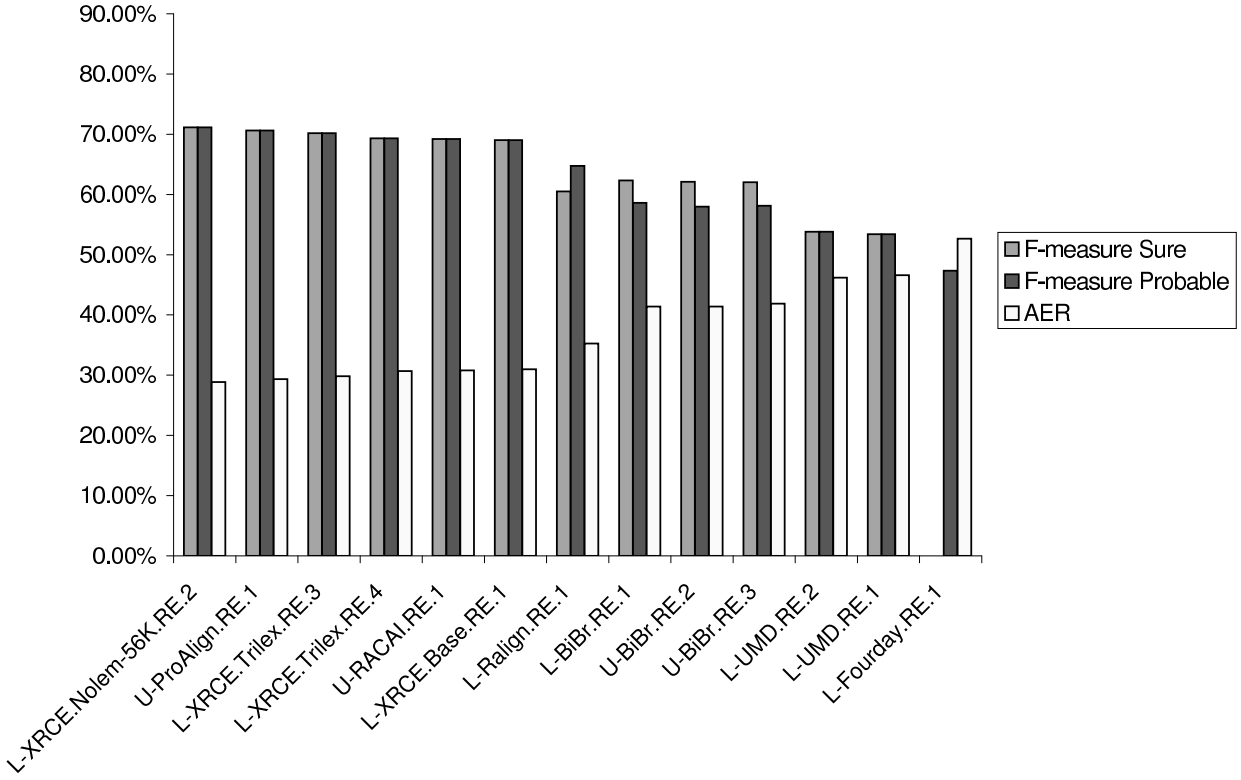

Figure 2: Ranked results for Romanian-English data

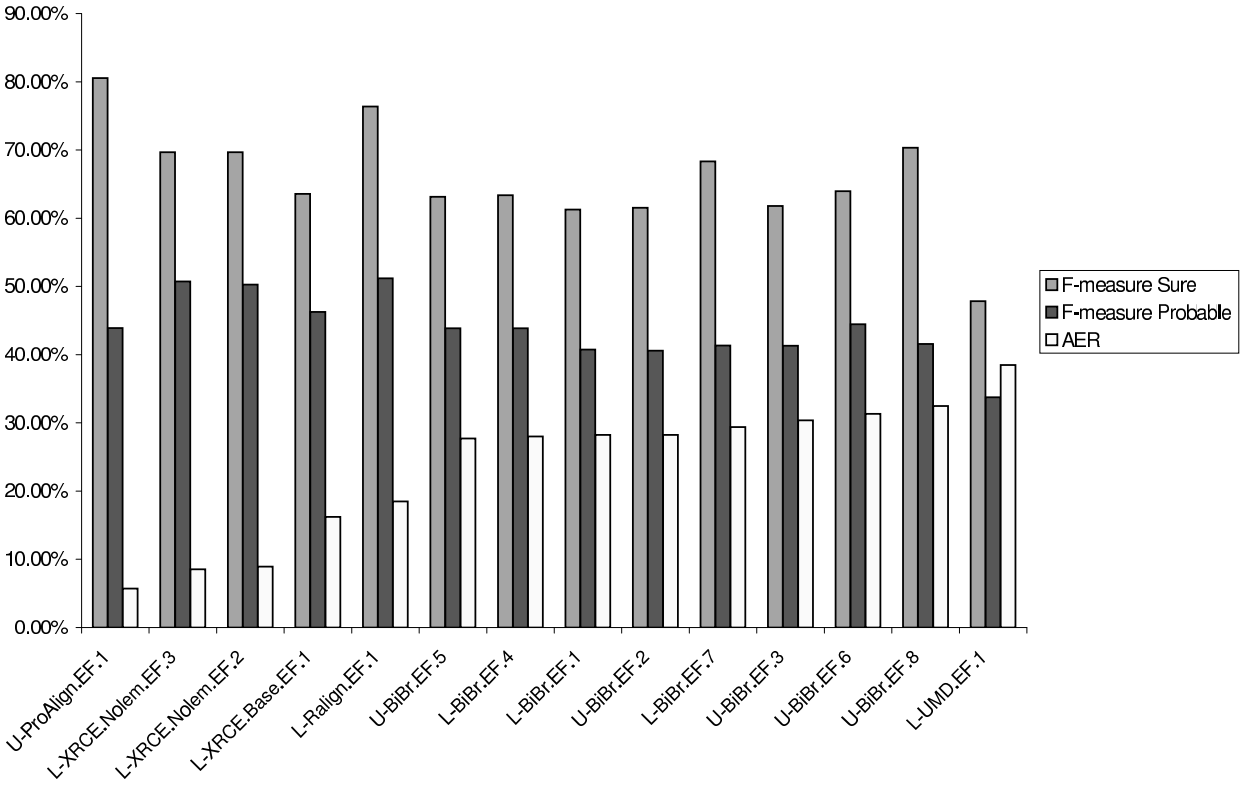

Figure 3: Ranked results for English-French data