ISI's Participation in the Romanian-English Alignment Task

Alexander Fraser
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292-6601
fraser@isi.edu

Daniel Marcu
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292-6601
marcu@isi.edu

Abstract

We discuss results on the shared task of Romanian-English word alignment. The baseline technique is that of symmetrizing two word alignments automatically generated using IBM Model 4. A simple vocabulary reduction technique results in an improvement in performance. We also report on a new alignment model and a new training algorithm based on alternating maximization of likelihood with minimization of error rate.

1 Introduction

ISI participated in the WPT05 Romanian-English word alignment task. The system used for baseline experiments is two runs of IBM Model 4 (Brown et al., 1993) in the GIZA++ (Och and Ney, 2003) implementation, which includes smoothing extensions to Model 4. For symmetrization, we found that Och and Ney's "refined" technique described in (Och and Ney, 2003) produced the best AER for this data set under all experimental conditions.

We experimented with a statistical model for inducing a stemmer cross-lingually, but found that the best performance was obtained by simply lowercasing both the English and Romanian text and removing all but the first four characters of each word.

We also tried a new model and a new training criterion based on alternating the maximization of likelihood and minimization of the alignment error rate. For these experiments, we have implemented an alignment package for IBM Model 4 using a hillclimbing search and Viterbi training as described in (Brown et al., 1993), and extended this to use new submodels. The starting point is the final alignment generated using GIZA++'s implementation of IBM Model 1 and the Aachen HMM model (Vogel et al., 1996).

Paper organization: Section 2 is on the baseline, Section 3 discusses vocabulary reduction, Section 4 introduces our new model and training method, Section 5 describes experiments, Section 6 concludes.

We use the following notation: e refers to an English sentence composed of English words labeled e_i. f refers to a Romanian sentence composed of Romanian words labeled f_j. a is an alignment of e to f. We use the term "Viterbi alignment" to denote the most probable alignment we can find, rather than the true Viterbi alignment.

2 Baseline

To train our systems, Model 4 was trained two times, first using Romanian as the source language and then using English as the source language. For each training, we ran 5 iterations of Model 1, 5 iterations of the HMM model and 3 iterations of Model 4. For the distortion calculations of Model 4, we removed the dependencies on Romanian and English word classes. We applied the "union", "intersection" and "refined" symmetrization metrics (Och and Ney, 2003) to the final alignments output from training, as well as evaluating the two final alignments directly.

We tried to have a strong baseline. GIZA++ has many free parameters which can not be estimated using Maximum Likelihood training. We did not use

Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 91–94,
Ann Arbor, June 2005. ©Association for Computational Linguistics, 2005

the defaults, but instead used settings which produce good AER results on French/English bitext. We also optimized $p0$ on the 2003 test set (using AER), rather than using likelihood training. Turning off the extensions to GIZA++ and training $p0$ as in (Brown et al., 1993) produces a substantial increase in AER.

3 Vocabulary Size Reduction

Romanian is a Romance language which has a system of suffixes for inflection which is richer than English. Given the small amount of training data, we decided that vocabulary size reduction was desirable. As a baseline for vocabulary reduction, we tried reducing words to prefixes of varying sizes for both English and Romanian after lowercasing the corpora. We also tried Porter stemming (Porter, 1997) for English.

(Rogati et al., 2003) extended Model 1 with an additional hidden variable to represent the split points in Arabic between the prefix, the stem and the suffix to generate a stemming for use in Cross-Lingual Information Retrieval. As in (Rogati et al., 2003), we can find the most probable stemming given the model, apply this stemming, and retrain our word alignment system. However, we can also use the modified model directly to find the best word alignment without converting the text to its stemmed form.

We introduce a variable r_j for the Romanian stem and a variable s_j for the Romanian suffix (which when concatenated together give us the Romanian word f_j) into the formula for the probability of generating a Romanian word f_j using an alignment a_j given only an English sentence e. We use the index z to denote a particular stemming possibility. For a given Romanian word the stemming possibilities are simply every possible split point where the stem is at least one character (this includes the null suffix).

$$p(f_j, a_j|e) = \sum_z p(r_{j,z}, s_{j,z}, a_j|e) \qquad (1)$$

If the assumption is made that the stem and the suffix are generated independently from e, we can assume conditional independence.

$$p(f_j, a_j|e) = \sum_z p(r_{j,z}, a_j|e)p(s_{j,z}, a_j|e) \qquad (2)$$

We performed two sets of experiments, one set where the English was stemmed using the Porter stemmer and one set where each English word was stemmed down to its first four characters. We tried the best performing scoring heuristic for Arabic from (Rogati et al., 2003) where $p(s_{j,z}, a_j|e)$ is modeled using the heuristic $p(s_{j,z}|l_j)$ where $s_{j,z}$ is the Romanian suffix, and l_j is the last letter of the Romanian word f_j; these adjustments are updated during EM training. We also tried several other approximations of $p(s_{j,z}, a_j|e)$ with and without updates in EM training. We were unable to produce better results and elected to use the baseline vocabulary reduction technique for the shared task.

4 New Model and Training Algorithm

Our motivation for a new model and a new training approach which combines likelihood maximization with error rate minimization is threefold:

- Maximum Likelihood training of Model 4 is not sufficient to find good alignments
- We would like to model factors not captured by IBM Model 4
- Using labeled data could help us produce better alignments, but we have very few labels

We create a new model and train it using an algorithm which has a step which increases likelihood (like one iteration in the EM algorithm), alternating with a step which decreases error. We accomplish this by:

- grouping the parameters of Model 4 into 5 submodels
- implementing 6 new submodels
- combining these into a single log-linear model with 11 weights, λ_1 to λ_{11}, which we group into the vector λ
- defining a search algorithm for finding the alignment of highest probability given the submodels and λ
- devising a method for finding a λ which minimizes alignment error given fixed submodels and a set of gold standard alignments
- inventing a training method for alternating steps which estimate the submodels by increasing likelihood with steps which set λ to decrease alignment error

The submodels in our new alignment model are listed in table 1, where for ease of exposition we

Table 1: Submodels used for alignment

1	$t(f_j\|e_i)$	translation probabilities
2	$n(\phi_i\|e_i)$	fertility probabilities, ϕ_i is the number of words generated by the English word e_i
3	$null$	parameters used in generating Romanian words from English NULL word (including $p0$, $p1$)
4	$d_1(\triangle j)$	movement (distortion) probabilities of first Romanian word generated from English word
5	$d_{>1}(\triangle j)$	movement (distortion) probabilities of other Romanian words generated from English word
6		ttable estimated from intersection of two starting alignments for this iteration
7		translation table from English to Romanian Model 1 Iteration 5
8		translation table from Romanian to English Model 1 Iteration 5
9		backoff fertility (fertility estimated over all English words)
10		zero fertility English word penalty
11		non-zero fertility English word penalty

consider English to be the source language and Romanian the target language.

The log-linear alignment model is specified by equation 3. The model assigns non-zero probabilities only to 1-to-many alignments, like Model 4. (Cettolo and Federico, 2004) used a log-linear model trained using error minimization for the translation task, 3 of the submodels were taken from Model 4 in a similar way to our first 5 submodels.

$$p_\lambda(a, f|e) = \frac{exp(\sum_m \lambda_m h_m(f, a, e))}{\sum_{f,e,a} exp(\sum_m \lambda_m h_m(f, a, e))} \quad (3)$$

Given λ, the alignment search problem is to find the alignment a of highest probability according to equation 3. We solve this using the local search defined in (Brown et al., 1993).

We set λ as follows. Given a sequence A of alignments we can calculate an error function, $E(A)$. For these experiments average sentence AER was used. We wish to minimize this error function, so we select λ accordingly:

$$\operatorname*{argmin}_{\lambda} \sum_{\tilde{a}} E(\tilde{a})\delta(\tilde{a}, (\operatorname*{argmax}_{a} p_\lambda(a, f|e))) \quad (4)$$

Maximizing performance for all of the weights at once is not computationally tractable, but (Och, 2003) has described an efficient one-dimensional search for a similar problem. We search over each λ_m (holding the others constant) using this technique to find the best λ_m to update and the best value to update it to. We repeat the process until no further gain can be found.

Our new training method is:

REPEAT

- Start with submodels and lambda from previous iteration
- Find Viterbi alignments on entire training corpus using new model (similar to E-step of Model 4 training)
- Reestimate submodel parameters from Viterbi alignments (similar to M-step of Model 4 Viterbi training)
- Find a setting for λ that reduces AER on discriminative training set (new D-step)

We use the first 148 sentences of the 2003 test set for the discriminative training set. 10 settings for λ are found, the hypothesis list is augmented using the results of 10 searches using these settings, and then another 10 settings for λ are found. We then select the best λ. The discriminative training regimen is otherwise similar to (Och, 2003).

5 Experiments

Table 2 provides a comparison of our baseline systems using the “refined” symmetrization metric with the best limited resources track system from WPT03 (Dejean et al., 2003) on the 2003 test set. The best results are obtained by stemming both English and Romanian words to the first four letters, as described in section 2.

Table 3 provides details on our shared task submission. RUN1 is the word-based baseline system. RUN2 is the stem-based baseline system. RUN4 uses only the first 6 submodels, while RUN5 uses all 11 submodels. RUN3 had errors in processing, so we omit it.

Results:

- Our new 1-to-many alignment model and training method are successful, producing decreases of 0.03 AER when the source is Romanian, and 0.01 AER when the source is English.

Table 2: Summary of results for 2003 test set

System	Stem Sizes	AER
Xerox “nolem-er-56k”		0.289
Baseline	No processing	0.284
Baseline	Eng Porter / Rom 4	0.251
Baseline	Eng 4 / Rom 4	0.248

Table 3: Full results on shared task submissions (blind test 2005)

Run Names	Stem Sizes	Source Rom	Source Eng	Union	Intersection	Refined
ISI.RUN1	No processing	0.3834	0.3589	0.3590	0.3891	0.3165
ISI.RUN2	Eng 4 / Rom 4	0.3056	0.2880	0.2912	0.3041	0.2675
ISI.RUN4	Eng 4 / Rom 4	0.2798	0.2833	0.2773	0.2862	0.2663
ISI.RUN5	Eng 4 / Rom 4	0.2761	0.2778	0.2736	0.2807	0.2655

- These decreases do not translate to a large improvement in the end-to-end task of producing many-to-many alignments with a balanced precision and recall. We had a very small decrease of 0.002 AER using the “refined” heuristic.
- The many-to-many alignments produced using “union” and the 1-to-1 alignments produced using “intersection” were also improved.
- It may be a problem that we trained $p0$ using likelihood (it is in submodel 3) rather than optimizing $p0$ discriminatively as we did for the baseline.

6 Conclusion

- Considering multiple stemming possibilities for each word seems important.
- Alternating between increasing likelihood and decreasing error rate is a useful training approach which can be used for many problems.
- Our model and training method improve upon a strong baseline for producing 1-to-many alignments.
- Our model and training method can be used with the “intersection” heuristic to produce higher quality 1-to-1 alignments
- Models which can directly model many-to-many alignments and do not require heuristic symmetrization are needed to produce higher quality many-to-many alignments. Our training method can be used to train them.

7 Acknowledgments

This work was supported by DARPA-ITO grant NN66001-00-1-9814 and NSF grant IIS-0326276.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Mauro Cettolo and Marcello Federico. 2004. Minimum error training of log-linear translation models. In *Proc. of the International Workshop on Spoken Language Translation*, pages 103–106, Kyoto, Japan.

Herve Dejean, Eric Gaussier, Cyril Goutte, and Kenji Yamada. 2003. Reducing parameter space for word alignment. In *HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Alberta, July.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

M. F. Porter. 1997. An algorithm for suffix stripping. In *Readings in information retrieval*, pages 313–316, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Monica Rogati, Scott McCarley, and Yiming Yang. 2003. Unsupervised learning of arabic stemming using a parallel corpus. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, July.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING ’96: The 16th Int. Conf. on Computational Linguistics*, pages 836–841, Copenhagen, Denmark, August.