An Information Theoretic Approach to Bilingual Word Clustering

Manaal Faruqui and **Chris Dyer**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
{mfaruqui, cdyer}@cs.cmu.edu

Abstract

We present an information theoretic objective for bilingual word clustering that incorporates both monolingual distributional evidence as well as cross-lingual evidence from parallel corpora to learn high quality word clusters jointly in any number of languages. The monolingual component of our objective is the average mutual information of clusters of adjacent words in each language, while the bilingual component is the average mutual information of the aligned clusters. To evaluate our method, we use the word clusters in an NER system and demonstrate a statistically significant improvement in F_1 score when using bilingual word clusters instead of monolingual clusters.

1 Introduction

A word cluster is a group of words which ideally captures syntactic, semantic, and distributional regularities among the words belonging to the group. Word clustering is widely used to reduce the number of parameters in statistical models which leads to improved generalization (Brown et al., 1992; Kneser and Ney, 1993; Clark, 2003; Koo et al., 2008; Turian et al., 2010), and multilingual clustering has been proposed as a means to improve modeling of translational correspondences and to facilitate projection of linguistic resource across languages (Och, 1999; Täckström et al., 2012). In this paper, we argue that generally more informative clusters can be learned when evidence from multiple languages is considered while creating the clusters.

We propose a novel bilingual word clustering objective (§2). The first term deals with each language independently and ensures that the data is well-explained by the clustering in a sequence model (§2.1). The second term ensures that the cluster alignments induced by a word alignment have high mutual information across languages (§2.2). Since the objective consists of terms representing the entropy monolingual data (for each language) and parallel bilingual data, it is particularly attractive for the usual situation in which there is much more monolingual data available than parallel data. Because of its similarity to the variation of information metric (Meilă, 2003), we call this bilingual term in the objective the **aligned variation of information**.

2 Word Clustering

A word clustering $\mathcal{C}$ is a partition of a vocabulary $\Sigma = \{x_1, x_2, \ldots, x_{|\Sigma|}\}$ into K disjoint subsets, $C_1, C_2, \ldots, C_K$. That is, $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$; $C_i \cap C_j = \emptyset$ for all $i \neq j$ and $\bigcup_{k=1}^{K} C_k = \Sigma$.

2.1 Monolingual objective

We use the average surprisal in a probabilistic sequence model to define the monolingual clustering objective. Let c_i denote the word class of word w_i. Our objective assumes that the probability of a word sequence $\mathbf{w} = \langle w_1, w_2, \ldots, w_M \rangle$ is

$$p(\mathbf{w}) = \prod_{i=1}^{M} p(c_i \mid c_{i-1}) \times p(w_i \mid c_i), \qquad (2.1)$$

where c_0 is a special start symbol. The term $p(c_i \mid c_{i-1})$ is the probability of class c_i following class c_{i-1}, and $p(w_i \mid c_i)$ is the probability of class c_i emitting word w_i. Using the MLE esitmates after taking the negative logarithm, this term reduces to

the following as shown in (Brown et al., 1992):

$$H(\mathcal{C};\mathbf{w}) = 2\sum_{k=1}^{K}\frac{\#(C_k)}{M}\log\frac{\#(C_k)}{M} - \sum_{i}\sum_{j\neq i}\frac{\#(C_i,C_j)}{M}\log\frac{\#(C_i,C_j)}{M}$$

where $\#(C_k)$ is the count of C_k in the corpus $\mathbf{w}$ under the clustering $\mathcal{C}$, $\#(C_i, C_j)$ is the count of the number of times that cluster C_i precedes C_j and M is the size of the corpus. Using the monolingual objective to cluster, we solve the following search problem:

$$\hat{\mathcal{C}} = \arg\min_{\mathcal{C}} H(\mathcal{C};\mathbf{w}). \tag{2.2}$$

2.2 Bilingual objective

Now let us suppose we have a second language with vocabulary $\Omega = \{y_1, y_2, \ldots, y_{|\Omega|}\}$, which is clustered into K disjoint subsets $\mathcal{D} = \{D_1, D_2, \ldots, D_K\}$, and a corpus of text in the second language, $\mathbf{v} = \langle v_1, v_2, \ldots, v_N\rangle$. Obviously we can cluster both languages using the monolingual objective above:

$$\hat{\mathcal{C}}, \hat{\mathcal{D}} = \arg\min_{\mathcal{C},\mathcal{D}} H(\mathcal{C};\mathbf{w}) + H(\mathcal{D};\mathbf{v}).$$

This joint minimization for the clusterings for both languages clearly has no benefit since the two terms of the objective are independent. We must alter the object by further assuming that we have *a priori* beliefs that some of the words in $\mathbf{w}$ and $\mathbf{v}$ have the same meaning.

To encode this belief, we introduce the notion of a **weighted vocabulary alignment** $\mathcal{A}$, which is a function on pairs of words in vocabularies Σ and Ω to a value greater than or equal to 0, i.e., $\mathcal{A} : \Sigma \times \Omega \mapsto \mathbb{R}_{\geq 0}$. For concreteness, $\mathcal{A}(x, y)$ will be the number of times that x is aligned to y in a word aligned parallel corpus. By abuse of notation, we write marginal weights $\mathcal{A}(x) = \sum_{y\in\Omega}\mathcal{A}(x,y)$ and $\mathcal{A}(y) = \sum_{x\in\Sigma}\mathcal{A}(x,y)$. We also define the set marginals $\mathcal{A}(C,D) = \sum_{x\in C}\sum_{y\in D}\mathcal{A}(x,y)$.

Using this weighted vocabulary alignment, we state an objective that encourages clusterings to have high average mutual information when alignment links are followed; that is, on average how much information does knowing the cluster of a word $x \in \Sigma$ impart about the clustering of $y \in \Omega$, and vice-versa?

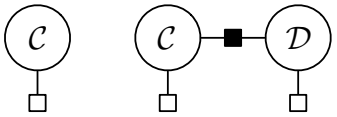

Figure 1: Factor graphs of the monolingual (left) & proposed bilingual clustering problem (right).

We call this quantity the **aligned variation of information** (AVI).

$$\text{AVI}(\mathcal{C},\mathcal{D};\mathcal{A}) = \mathbb{E}_{\mathcal{A}(x,y)}\left[-\log p(c_x \mid d_y) - \log p(d_y \mid c_x)\right]$$

Writing out the expectation and gathering terms, we obtain

$$\text{AVI}(\mathcal{C},\mathcal{D};\mathcal{A}) = -\sum_{x\in\Sigma}\sum_{y\in\Omega}\frac{\mathcal{A}(x,y)}{\mathcal{A}(\cdot,\cdot)}\times \left[2\log\frac{\mathcal{A}(C,D)}{\mathcal{A}(\cdot,\cdot)} - \log p(C) - \log p(D)\right],$$

where it is assumed that $0 \log x = 0$.

Our bilingual clustering objective can therefore be stated as the following search problem over a linear combination of the monolingual and bilingual objectives:

$$\arg\min_{\mathcal{C},\mathcal{D}} \overbrace{H(\mathcal{C};\mathbf{w}) + H(\mathcal{D};\mathbf{v})}^{\text{monolingual}} + \overbrace{\beta\text{AVI}(\mathcal{C},\mathcal{D})}^{\beta\times\text{bilingual}}. \tag{2.3}$$

Understanding AVI. Intuitively, we can imagine sampling a random alignment from the distribution obtained by normalizing $\mathcal{A}(\cdot,\cdot)$. AVI gives us a measure of how much information do we obtain, on average, from knowing the cluster in one language about the clustering of a linked element chosen at random proportional to $\mathcal{A}(x,\cdot)$ (or conditioned the other way around). In the following sections, we denote $\text{AVI}(\mathcal{C},\mathcal{D};\mathcal{A})$ by $\text{AVI}(\mathcal{C},\mathcal{D})$. To further understand AVI, we remark that AVI reduces to the VI metric when the alignment maps words to themselves in the same language. As a proper metric, VI has a number of attractive properties, and these can be generalized to AVI (without restriction on the alignment map), namely:

- *Non-negativity:* $\text{AVI}(C,D) \geq 0$;
- *Symmetry:* $\text{AVI}(C,D) = \text{AVI}(D,C)$;
- *Triangle inequality:* $\text{AVI}(C,D) + \text{AVI}(D,E) \geq \text{AVI}(C,E)$;

- *Identity of indiscernables:* $\text{AVI}(C, D) = 0$ iff $C \equiv D$.[1]

2.3 Example

Figure 2 provides an example illustrating the difference between the bilingual vs. monolingual clustering objectives. We compare two different clusterings of a two-sentence Arabic-English parallel corpus (the English half of the corpus contains the same sentence, twice, while the Arabic half has two variants with the same meaning). While English has a relatively rigid SVO word order, Arabic can alternate between the traditional VSO order and an more modern SVO order. Since our monolingual clustering objective relies exclusively on the distribution of clusters before and after each token, flexible word order alternations like this can cause unintuitive results. To further complicate matters, verbs can inflect differently depending on whether their subject precedes or follows them (Haywood and Nahmad, 1999), so a monolingual model, which knows nothing about morphology and may only rely on distributional clues, has little chance of performing well without help. This is indeed what we observe in the monolingual objective optimal solution (center), in which `AwlAd` (*boys*) and `yElbwn` (*play*+PRES + 3PL) are grouped into a single class, while `yElb` (*play*+PRES + 3SG) is in its own class. However, the AVI term (which is of course not included) has a value of 1.0, reflecting the relatively disordered clustering relative to the given alignment. On the right, we see the optimal solution that includes the AVI term in the clustering objective. This has an AVI of 0, indicating that knowing the clustering of any word is completely informative about the words it is aligned to. By including this term, a slightly worse monolingual solution is chosen, but the clustering corresponds to the reasonable intuition that words with the same meaning (i.e., the two variants of *to play*) should be clustered together.

2.4 Inference

Figure 1 shows the factor graph representation of our clustering models. Finding the optimal clustering under both the monolingual and bilingual objectives is a computationally hard combinatorial optimization problem (Och, 1995). We use a greedy hill-climbing word exchange algorithm (Martin et al., 1995) to find a minimum value for our objective. We terminate the optimization procedure when the number of words exchanged at the end of one complete iteration through both the languages is less than 0.1% of the sum of vocabulary of the two languages and at least five complete iterations have been completed.[2] For every language the word clusters are initialised in a round robin order according to the token frequency.

[1] $C \equiv D$ iff $\forall i |\{\mathcal{D}(y)|\forall(x, y) \in \mathcal{A}, \mathcal{C}(x) = i\}| = 1$

3 Experiments

Evaluation of clustering is not a trivial problem. One branch of work seeks to recast the problem as the of part-of-speech (POS) induction and attempts to match linguistic intuitions. However, hard clusters are particularly useful for downstream tasks (Turian et al., 2010). We therefore chose to focus our evaluation on the latter problem. For our evaluation, we use our word clusters as an input to a named entity recognizer which uses these clusters as a source of features. Our evaluation task is the German corpus with NER annotation that was created for the shared task at CoNLL-2003 [3]. The training set contains approximately 220,000 tokens and the development set and test set contains 55,000 tokens each. We use Stanford's Named Entity Recognition system[4] which uses a linear-chain conditional random field to predict the most likely sequence of NE labels (Finkel and Manning, 2009).

Corpora for Clustering: We used parallel corpora for {Arabic, English, French, Korean & Turkish}-German pairs from WIT-3 corpus (Cettolo et al., 2012) [5], which is a collection of translated transcriptions of TED talks. Each language pair contained around 1.5 million German words. The corpus was word aligned in two directions using an unsupervised word aligner (Dyer et al., 2013), then the intersected alignment points were taken.

Monolingual Clustering: For every language pair, we train German word clusters on the monolingual German data from the parallel data. Note that the parallel corpora are of different sizes and hence the monolingual German data from every parallel corpus is different. We treat the F_1 score

[2] In practice, the number of exchanged words drops of exponentially, so this threshold is typically reached in not many iterations.

[3] http://www.cnts.ua.ac.be/conll2003/ner/

[4] http://nlp.stanford.edu/ner/index.shtml

[5] https://wit3.fbk.eu/mt.php?release=2012-03

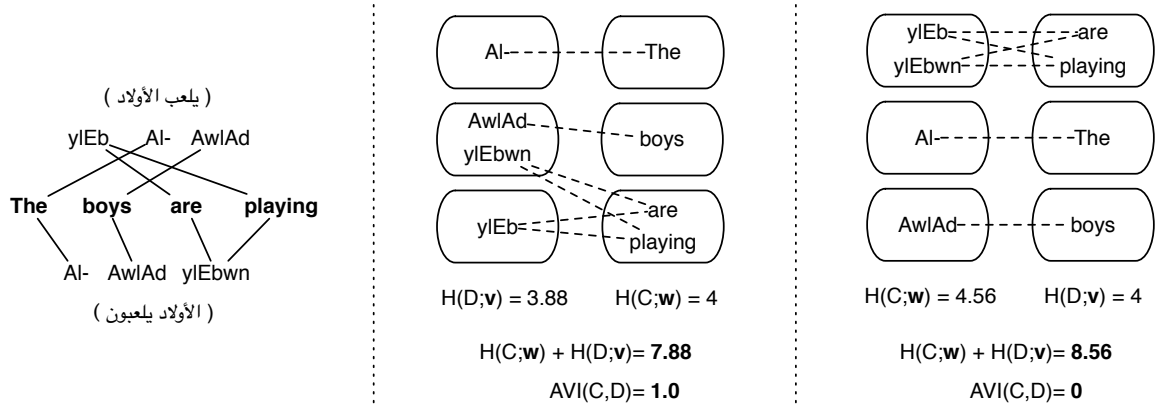

Figure 2: A two-sentence English-Arabic parallel corpus (left); a 3-class clustering that maximizes the monolingual objective ($\beta = 0$; center); and a 3-class clustering that maximizes the joint monolingual and bilingual objective (any $\beta > 0.68$; right).

obtained using monolingual word clusters ($\beta = 0$) as the baseline. Table 1 shows the F_1 score of NER[6] when trained on these monolingual German word clusters.

Bilingual Clustering: While we have formulated a joint objective that enables using both monolingual and bilingual evidence, it is possible to create word clusters using the bilingual signal only by removing the first term in Eq. 2.3. Table 1 shows the performance of NER when the word clusters are obtained using only the bilingual information for different language pairs. As can be seen, these clusters are helpful for all the language pairs. For *Turkish* the F_1 score improves by 1.0 point over when there are no distributional clusters which clearly shows that the word alignment information improves the clustering quality. We now need to supplement the bilingual information with monolingual information to see if the improvement sustains.

We varied the weight of the bilingual objective (β) from 0.05 to 0.9 and observed the effect in NER performance on English-German language pair. The F_1 score is maximum for $\beta = 0.1$ and decreases monotonically when β is either increased or decreased. This indicates that bilingual information is helpful, but less valuable than monolingual information. Preliminary experiments showed that the value of $\beta = 0.1$ is fairly robust across other language pairs and hence we fix it to that for all the experiments.

We run our bilingual clustering model ($\beta = 0.1$) across all language pairs and note the F_1 scores. Table 1 (unrefined) shows that except for Arabic-German & French-German, all other language pairs deliver a better F_1 score than only using monolingual German data. In case of Arabic-German there is a drop in score by 0.25 points. Although, we have observed improvement in F_1 score over the monolingual case, the gains do not reach significance according to McNemar's test (Dietterich, 1998).

Thus we propose to further refine the quality of word alignment links as follows: Let x be a word in language Σ and y be a word in language Ω and let there exists an alignment link between x and y. Recall that $\mathcal{A}(x, y)$ is the count of the alignment links between x and y observed in the parallel data, and $\mathcal{A}(x)$ and $\mathcal{A}(y)$ are the respective marginal counts. Then we define an edge association weight $e(x, y) = \frac{2\times\mathcal{A}(x,y)}{\mathcal{A}(x)+\mathcal{A}(y)}$ This quantity is an association of the strength of the relationship between x and y, and we use it to remove all alignment links whose $e(x, y)$ is below a given threshold before running the bilingual clustering model. We vary e from 0.1 to 0.7 and observe the new F_1 scores on the development data. Table 1 (refined) shows the results obtained by our refined model. The values shown in bold are the highest improvements over the monolingual model.

For English and Turkish we observe a statistically significant improvement over the monolingual model (cf. Table 1) with $p < 0.007$ and $p < 0.001$ according to McNemar's test. Arabic improves least with just an improvement of 0.02 F_1 points over the monolingual baseline. We

[6] Faruqui and Padó (2010) show that for the size of our generalization data in German-NER, $K = 100$ should give us the optimum value.

Language Pair	Dev				Test	
	— (only bi)	$\beta = 0$ (only mono)	$\beta = 0.1$ (unrefined)	$\beta = 0.1$ (refined)	$\beta = 0$ (only mono)	$\beta = 0.1$ (refined)
No clusters	68.27				72.32	
En-De	68.95	70.04	**70.33**	**70.64**†	72.30	**72.98**†
Fr-De	69.16	69.74	69.69	**69.89**	72.66	**72.83**
Ar-De	69.01	69.65	69.40	**69.67**	72.90	72.37
Tr-De	69.29	69.46	**69.64**	**70.05**†	72.41	**72.54**
Ko-De	68.95	69.70	**69.78**	**69.95**	72.71	72.54
Average	69.07	69.71	**69.76**	**70.04**†	72.59	**72.65**

Table 1: NER performance using different word clustering models. Bold indicates an improvement over the monolingual ($\beta = 0$) baseline; † indicates a significant improvement (McNemar's test, $p < 0.01$).

see that the optimal value of e changes from one language pair to another. For French and English $e = 0.1$ gives the best results whereas for Turkish and Arabic $e = 0.5$ and for Korean $e = 0.7$. Are these thresholds correlated with anything? We suggest that higher values of e correspond to more intrinsically noisy alignments. Since alignment models are parameterized based on the vocabularies of the languages they are aligning, larger vocabularies are more prone to degenerate solutions resulting from overfitting. So we are not surprised to see that sparser alignments (resulting from higher values of e) are required by languages like Korean, while languages like French and English make due with denser alignments.

Evaluation on Test Set: We now verify our results on the test set. We take the best bilingual word clustering model obtained for every language pair ($e = 0.1$ for En, Fr. $e = 0.5$ for Ar, Tr. $e = 0.7$ for Ko) and train NER classifiers using these. Table 1 shows the performance of German NER classifiers on the test set. All the values shown in bold are better than the monolingual baselines. English again has a statistically significant improvement over the baseline. French and Turkish show the next best improvements. The English-German cluster model performs better than the `mkcls`[7] tool (72.83%).

4 Related Work

Our monolingual clustering model is purely distributional in nature. Other extensions to word clustering have incorporated morphological and orthographic information (Clark, 2003). The work of Snyder and Barzilay (2010), which focused on POS induction is very closely related. The earliest work on bilingual word clustering was proposed by (Och, 1999) which, like us, uses a language modeling approach (Brown et al., 1992; Kneser and Ney, 1993) for monolingual optimization and a similarity function for bilingual similarity. Täckström et al. (2012) use cross-lingual word clusters to show transfer of linguistic structure. While their clustering method is superficially similar, the objective function is more heuristic in nature than our information-theoretic conception of the problem. Multilingual learning has been applied to a number of unsupervised and supervised learning problems, including word sense disambiguation (Diab, 2003; Guo and Diab, 2010), topic modeling (Mimno et al., 2009; Boyd-Graber and Blei, 2009), and morphological segmentation (Snyder and Barzilay, 2008).

Also closely related is the technique of cross-lingual annotation projection. This has been applied to bootstrapping syntactic parsers (Hwa et al., 2005; Smith and Smith, 2007; Cohen et al., 2011), morphology (Fraser, 2009), tense (Schiehlen, 1998) and T/V pronoun usage (Faruqui and Padó, 2012).

5 Conclusions

We presented a novel information theoretic model for bilingual word clustering which seeks a clustering with high average mutual information between clusters of adjacent words, and also high mutual information across observed word alignment links. We have shown that improvement in clustering can be obtained across a range of language pairs, evaluated in terms of their value as features in an extrinsic NER task. Our model can be extended for clustering any number of given languages together in a joint framework, and incorporate both monolingual and parallel data.

Acknowledgement: We woud like to thank W. Ammar, V. Chahuneau and W. Ling for valuable discussions.

[7] http://www.statmt.org/moses/giza/mkcls.html

References

J. Boyd-Graber and D. M. Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 75–82, Arlington, Virginia, United States. AUAI Press.

P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. 1992. Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–479, December.

M. Cettolo, C. Girardi, and M. Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.

A. Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 59–66, Stroudsburg, PA, USA. Association for Computational Linguistics.

S. B. Cohen, D. Das, and N. A. Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 50–61, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. T. Diab. 2003. *Word sense disambiguation within a multilingual framework*. Ph.D. thesis, University of Maryland at College Park, College Park, MD, USA. AAI3115805.

T. G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923.

C. Dyer, V. Chahuneau, and N. A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proc. NAACL*.

M. Faruqui and S. Padó. 2010. Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany.

M. Faruqui and S. Padó. 2012. Towards a model of formal and informal address in english. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

J. R. Finkel and C. D. Manning. 2009. Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 141–150, Stroudsburg, PA, USA. Association for Computational Linguistics.

A. Fraser. 2009. Experiments in morphosyntactic processing for translating to and from German. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 115–119, Athens, Greece, March. Association for Computational Linguistics.

W. Guo and M. Diab. 2010. Combining orthogonal monolingual and multilingual sources of evidence for all words wsd. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1542–1551, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. A. Haywood and H. M. Nahmad. 1999. *A new Arabic grammar of the written language*. Lund Humphries Publishers.

R. Hwa, P. Resnik, A. Weinberg, C. I. Cabezas, and O. Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, pages 311–325.

R. Kneser and H. Ney. 1993. Forming word classes by statistical clustering for statistical language modelling. In R. Khler and B. Rieger, editors, *Contributions to Quantitative Linguistics*, pages 221–226. Springer Netherlands.

T. Koo, X. Carreras, and M. Collins. 2008. Simple semi-supervised dependency parsing. In *Proc. of ACL*.

S. Martin, J. Liermann, and H. Ney. 1995. Algorithms for bigram and trigram word clustering. In *Speech Communication*, pages 1253–1256.

M. Meilă. 2003. Comparing Clusterings by the Variation of Information. In *Learning Theory and Kernel Machines*, pages 173–187.

D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 880–889, Stroudsburg, PA, USA. Association for Computational Linguistics.

F. J. Och. 1995. Maximum-Likelihood-Schätzung von Wortkategorien mit Verfahren der kombinatorischen Optimierung. Studienarbeit, University of Erlangen.

F. J. Och. 1999. An efficient method for determining bilingual word classes. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, pages 71–76, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Schiehlen. 1998. Learning tense translation from bilingual corpora.

D. A. Smith and N. A. Smith. 2007. Probabilistic Models of Nonprojective Dependency Trees. In *Proceedings of the 2007 Joint Conference on*

Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 132–140, Prague, Czech Republic, June. Association for Computational Linguistics.

B. Snyder and R. Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *In The Annual Conference of the Association for Computational Linguistics*.

B. Snyder and R. Barzilay. 2010. Climbing the tower of babel: Unsupervised multilingual learning. In J. Frnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 29–36. Omnipress.

O. Täckström, R. McDonald, and J. Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, page 11. Association for Computational Linguistics.

J. Turian, L. Ratinov, and Y. Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394, Stroudsburg, PA, USA. Association for Computational Linguistics.