Learning to translate from graded and negative relevance information

Laura Jehl
Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
jehl@cl.uni-heidelberg.de

Stefan Riezler
Computational Linguistics & IWR
Heidelberg University
69120 Heidelberg, Germany
riezler@cl.uni-heidelberg.de

Abstract

We present an approach for learning to translate by exploiting cross-lingual link structure in multilingual document collections. We propose a new learning objective based on structured ramp loss, which learns from graded relevance, explicitly including negative relevance information. Our results on English-German translation of Wikipedia entries show small, but significant, improvements of our method over an unadapted baseline, even when only a weak relevance signal is used. We also compare our method to monolingual language model adaptation and automatic pseudo-parallel data extraction and find small improvements even over these strong baselines.

1 Introduction

Typically, parameters of an SMT system are learned on a small parallel data set from the domain or genre of interest. However, while many multilingual data sets, especially in the realm of user-generated data, contain document-level links, sentence-parallel training data are not always available. A small number of sentences can be manually translated for in-domain parameter tuning, but this ignores most of the available multilingual resource. Monolingual language model adaptation via concatenation or interpolation is one viable solution which makes use of the target side part of a collection (see e.g. Koehn and Schroeder (2007) or Foster and Kuhn (2007)). Additionally, there are several approaches to automatic parallel data extraction from cross-lingual document-level links, such as Munteanu and Marcu (2005)'s work on news data, or more recent work on Wikipedia by Wołk and Marasek (2015), and on websites by Smith et al. (2013). We argue that these approaches work well if the cross-lingual links are a strong signal for parallelism, but fail if the signal linking documents across languages is weaker. We propose a method for tuning sparse lexicalized features on large amounts of multilingual data which contain some cross-lingual document-level relevance annotation. We do so by re-formulating the structured ramp loss objective proposed by Chiang (2012) and Gimpel and Smith (2012) to incorporate graded and negative cross-lingual relevance signals. Using translation of Wikipedia entries as a running example, we evaluate the efficacy of our method along with the traditional approaches on a manually created in-domain test set. We show that our method is able to produce small, but significant, gains, even if only a weak relevance signal is used.

Section 2 explains our learning objective and cost function. In Section 3 we describe the construction of our training and evaluation data, including pseudo-parallel data extraction. Section 4 contains details of our experimental setup and presents our experimental results. Section 5 concludes the paper.

2 Learning from graded relevance feedback

2.1 Learning objectives

We work within a scenario where we want to learn the parameters of an SMT system, but have no in-domain reference translations available. What we have, is a large collection of source and target language documents and a signal telling us that some target documents are *more relevant* to a source document

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3156–3166, Osaka, Japan, December 11-17 2016.

than others. For example, in the multilingual Wikipedia, cross-lingual documents can be connected in different ways. First, a link exists between two documents if they are connected by an interlanguage link (we call this a *mate* relation). This is a very strong relevance signal. Second, a more indirect link exists between a source language document and a target language document if the target language document is connected to the source language document's mate by a hyperlink (we call this a *link* relation). This is a weaker relevance signal. A cross-lingual mate is more relevant to an input document than a document that is only linked to by the mate. In turn, this linked document is more relevant to the input than a document that has no direct link to the mate. Any Wikipedia document is more relevant than a document from a different data set. We write relevance as $d_1 \succ_{\mathbf{f}} d_2$ ("d_1 is more relevant to $\mathbf{f}$ than d_2"). Another example of graded relevance information, which has been used in information retrieval, occurs in multilingual patent collections, where patent documents can be in a "family" relation if they contain publications of the same patent, or be related to a lesser degree, if a target document is cited by a source document's family patent. Of course, other notions of relevance are conceivable, e.g. by document similarity, and we plan to further investigate such notions in future work.

In order to incorporate graded relevance information, we modify the structured ramp loss objective by Gimpel and Smith (2012) to also include negative relevance information. Ramp loss based SMT tuning methods as presented by Gimpel and Smith (2012) and Chiang (2012) usually try to find parameters that separate a "good" hypothesis with respect to the reference from one that is "bad" with respect to the same reference. Goodness and badness are measured by an external cost function, or a cost function combined with the model prediction. Equation 1 shows one version of the structured ramp loss ("ramp loss 3"/equation 8 from Gimpel and Smith (2012)):

$$L_{Gimpel}(\mathbf{F};\theta) = \sum_{\mathbf{f}\in\mathbf{F}} \underbrace{-\max_{\mathbf{e}}(\text{score}(\mathbf{e};\theta) - \text{cost}(\mathbf{e}))}_{\text{hope derivation}} + \underbrace{\max_{\mathbf{e}}(\text{score}(\mathbf{e};\theta) + \text{cost}(\mathbf{e}))}_{\text{fear derivation}} \quad (1)$$

where $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2 \ldots \mathbf{f}_n\}$ is a finite set of input examples, θ refers to the model parameters and $\mathbf{e}$ is a translation hypothesis; $\text{score}(\mathbf{e};\theta)$ is the log-linear model score of the hypothesis, which is proportional to the dot product between the feature vector associated with the hypothesis and the weight vector; $\text{cost}(\mathbf{e})$ is a cost function, which measures the quality of the current hypothesis. Usually, this function is some per-sentence approximation of the BLEU score against one or more reference translations. Following Chiang (2012)'s terminology, this loss tries to maximize the distance between a *hope* derivation – which has high model score and low cost – from a *fear* derivation – which has high model score, but high cost.

We define two training objectives which are variations of this loss function, but which incorporate positive and negative relevance information. Our intuition is that instead of trying to separate *hope* and *fear* with respect to the same reference, we try to separate a hypothesis that has high model score and low cost with respect to a relevant document from one that has high model score and low cost with respect to a document that is irrelevant. Our first objective is given in Equation 2:

$$L_{ramp_1}(\mathbf{F};\theta) = \sum_{\mathbf{f}\in\mathbf{F}} \underbrace{-\max_{\mathbf{e}}(\text{score}(\mathbf{e},\mathbf{f};\theta) - \text{cost}(\mathbf{e},d^+_{\mathbf{f}}))}_{\text{hope derivation w.r.t. } d^+} + \underbrace{\max_{\mathbf{e}}(\text{score}(\mathbf{e},\mathbf{f};\theta) - \text{cost}(\mathbf{e},d^-_{\mathbf{f}}))}_{\text{hope derivation w.r.t. } d^-} \quad (2)$$

In this objective, $\text{cost}(\mathbf{e},d) \in [0,1]$ is the cost of a hypothesis $\mathbf{e}$ with respect to a document d. d^+ and d^- are documents, such that $d^+ \succ_{\mathbf{f}} d^-$. Unlike L_{Gimpel}, this loss tries to separate two different hope derivations. One potential weakness of L_{ramp_1} is that it treats d^+ and d^- completely independently. This could lead to very similar hypotheses being selected, if there exist hypotheses that have low cost in both d^+ and in d^-. To solve this issue, we propose a second modification of the loss.

In the second variant we define "good" and "bad" hypotheses as those which have the largest *difference* between the cost with respect to d^+ and d^-, i.e. hypotheses that best distinguish d^+ from d^-. This leads to the following objective given in Equation 3.

$$L_{ramp_2}(\mathbf{F};\theta) = \sum_{\mathbf{f}\in\mathbf{F}} \underbrace{-\max_{\mathbf{e}}(\text{score}(\mathbf{e},\mathbf{f};\theta) - (\text{cost\_diff}(\mathbf{e},d^+_{\mathbf{f}},d^-_{\mathbf{f}}))}_{\text{derivation with lowest cost}(d^+)\textit{ and }\text{highest cost}(d^-)} + \underbrace{\max_{\mathbf{e}}(\text{score}(\mathbf{e},\mathbf{f};\theta) - (\text{cost\_diff}(\mathbf{e},d^-_{\mathbf{f}},d^+_{\mathbf{f}}))}_{\text{derivation with lowest cost}(d^-)\textit{ and }\text{highest cost}(d^+)} \quad (3)$$

where cost_diff is defined as

$$\text{cost\_diff}(\mathbf{e},d_1,d_2) = \text{cost}(\mathbf{e},d_1) - \text{cost}(\mathbf{e},d_2) \quad (4)$$

Note that with the above definition of cost_diff, equation 3 can be reformulated as

$$L_{ramp_2}(\mathbf{F};\theta) = \sum_{\mathbf{f}\in\mathbf{F}} \underbrace{-\max_{\mathbf{e}}(\text{score}(\mathbf{e},\mathbf{f};\theta) - \text{cost\_diff}(\mathbf{e},d^+_{\mathbf{f}},d^-_{\mathbf{f}}))}_{\text{hope derivation}} + \underbrace{\max_{\mathbf{e}}(\text{score}(\mathbf{e},\mathbf{f};\theta) + \text{cost\_diff}(\mathbf{e},d^+_{\mathbf{f}},d^-_{\mathbf{f}}))}_{\text{fear derivation}} \quad (5)$$

which is identical to the original structured ramp loss (Equation 1), but still allows to include positive and negative relevance signals via the cost function. We apply a linear scaling operation to squash our new cost function to return values between 0 and 1.

2.2 Implementation and learning

Parallelized stochastic subgradient descent. Algorithm 1 shows our learning procedure. Optimization is done using stochastic subgradient descent (SSD) as proposed for ramp loss by Keshet and McAllester (2011). In order to be able to train on thousands of documents, we use the method described in Algorithm 4 of Simianer et al. (2012), which splits training data into shards (line 1 in Algorithm 1), trains one epoch on each shard (line 3 to 12), and then applies feature selection by ℓ_1/ℓ_2 regularization (line 13) before starting the next epoch.

Sampling. For each training example, we first sample a document pair (d^+, d^-) (line 6). The *sample()* procedure draws documents d^+ from a set of relevant documents D^+ and d^- from a set of "contrast documents", D^-, according to some cross-lingual relevance signal. In our experiments, we first use random sampling. We also try out a weighted sampling strategy, if the relevance signal is weaker. In this case, we want to sample a document more frequently from D^+, if it is more similar to the input document. We calculate cross-lingual document similarity by using document representations from bilingual word embeddings. The embeddings are learned from the aligned parallel training corpus using the Bilingual Skip-gram model of Luong et al. (2015).[1] Document representations are computed by averaging over all word representations in the document, weighted by the inverse document frequencies of the words. Cosine similarity is used to measure similarity between the current source document and the documents in D^+. We use weighted reservoir sampling (Efraimidis and Spirakis, 2006) to draw a document weighted by its similarity to the current source document. The contrast document d^- is drawn randomly from D^-, but is re-drawn if d^- is more similar to the input then d^+.

Search. In lines 7 and 8 we identify the "good" and "bad" hypotheses h^+ and h^- by running *search()*. Most tuning algorithms use k-best lists to approximate the search space over possible translation hypotheses. However, k-best lists cover only a very small portion of the possible hypothesis space and often contain very similar hypotheses. Since we may not have a strong enough signal to differentiate between those hypotheses, we also experiment with using the entire search space, which in hierarchical phrase-based translation can be represented by a packed hypothesis forest, or hypergraph. In this scenario, *search()* amounts to finding the Viterbi derivation after annotating the translation hypergraph

[1] github.com/lmthang/bivec

Algorithm 1 SSD

Require: input X, epochs T, initial weights w_0, cost function $cost$, document collection D^+, D^-, stepsize η, regularization strength C, number of shards S

1: $\{X_1 \dots X_S\} \leftarrow$ make_shards(S, X) ▷ Create shards for parallel training
2: **for** $t = 1$ to T **do**
3: **for** $s = 1$ to S parallel **do**
4: $w_{s,t-1}^{(0)} \leftarrow w_{t-1}$
5: **for** $i = 1$ to $|X_s|$ **do**
6: $(d^+, d^-) \leftarrow$ sample$(X_s^{(i)}, D^+, D^-)$ ▷ Sample relevant and irrelevant document
7: $(h^+, h^-) \leftarrow$ search$(X_s^{(i)}, w_{s,t-1}^{(i-1)}, cost, d^+, d^-)$ ▷ Find hope and fear
8: $w_{s,t-1}^{(i)} \leftarrow w_{s,t-1}^{(i-1)} + \eta(\phi(h^+) - \phi(h^-)) - \eta C(\frac{w_{s,t-1}^{(i-1)} - w_0}{|X|})$ ▷ Update weights
9: **end for**
10: $w_{s,t} \leftarrow w_{s,t-1}^{(|X|)}$
11: **end for**
12: $w_t \leftarrow$ select$(w_{1,t} \dots w_{S,t})$ ▷ Select features by ℓ_1/ℓ_2 regularization
13: **end for**

edges with the cost for each edge. This requires a cost function which decomposes over edges, as will be detailed in section 2.3. We run experiments both using a k-best lists and the full search space.

Finally, the weights are updated in line 9 by adding the negative subgradient multiplied by learning rate η and a regularization term which is obtained from adding $C\frac{1}{2|X|}\|(w - w_0)^2\|$ to the ramp loss objective.

2.3 Cost function

So far, we have not yet specified the cost function. Usually, $1 - psBLEU(\mathbf{e}, \mathbf{r})$ is used as a cost function, where $\mathbf{r}$ is a reference translation and $psBLEU$ is a per-sentence approximation of the $BLEU$ score. Since we do not have reference translations as feedback, we need to use a cost function that will evaluate the quality of a hypothesis with respect to a relevant document. Like BLEU we use average n-gram precision. Unlike BLEU, we cannot use reference length to control the length of the produced translation. Our solution is to use the source length, multiplied by the average source-target length ratio r which can be empirically determined on the training set.

For *k*-best training, where we can evaluate complete sentences, we use average *n*-gram precision:

$$\mathrm{nprec}(\mathbf{e}, \mathbf{f}, d) = \frac{1}{N}\sum_{n=1}^{N} \frac{\sum_{u_n} c_{u_n}(\mathbf{e}) \cdot \delta_{u_n}(d)}{\sum_{u_n} c_{u_n}(\mathbf{e})} \cdot min(1, \frac{r \cdot |\mathbf{e}|}{|\mathbf{f}|})$$

where N is the maximum *n*-gram size, u_n are *n*-grams present in $\mathbf{e}$, $c_{u_n}(\mathbf{e})$ counts the occurrences of u_n in $\mathbf{e}$ and $\delta_{u_n}(d)$ returns 1 if u_n is present in document d and 0 otherwise. The second term is the brevity penalty. As a cost function, this becomes $1 - \mathrm{nprec}$. BLEU uses the geometric mean to account for the exponentially decaying precision, as n increases. When calculating per-sentence BLEU, we might face the problem of zero-precision, as n increases. Since BLEU is measured over a corpus and not over a sentence, the case of zero-values was not taken into consideration. A common solution to this is count smoothing. We use the arithmetic instead of the geometric mean, since it avoids the problem of zeros, and will return the same ranking as the geometric mean.

When training on hypergraphs, we are facing the problem that *n*-gram precision is not edge-decomposable. For our hypergraph experiments, we tried the simplest possible approach, which is to compute nprec at edge level:

$$\mathrm{nprec}(\mathbf{e}, \mathbf{f}, d) = \sum_{\bar{e} \in \mathbf{e}} \mathrm{nprec}(\bar{e}, \bar{f}, d)$$

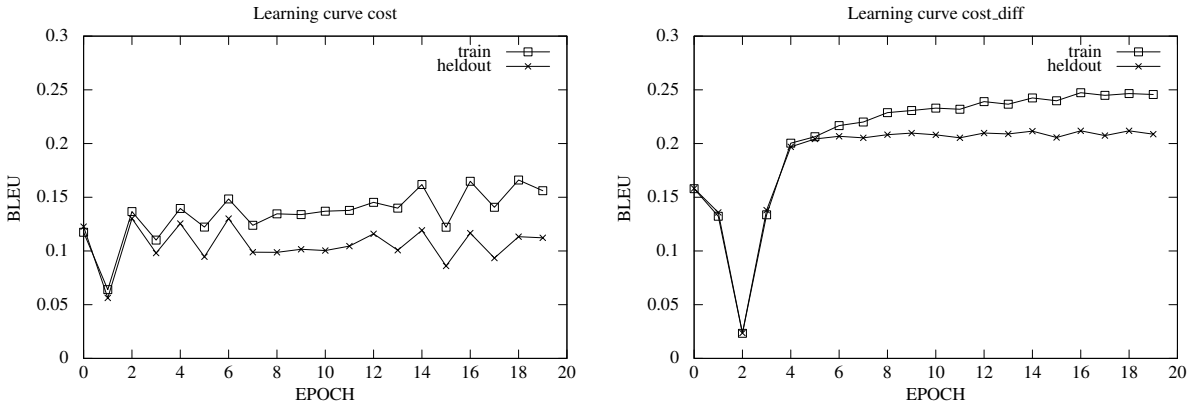

Figure 1: Learning curves on training and heldout data when training on references. The left side uses the loss from equation 2, the right side uses the loss from equation 3.

In order to test the proposed loss and cost functions, we look at how they perform in a case, where we are learning translation model weights from a perfect signal, i.e. reference translation. Instead of sampling d^+ from a set of relevant documents D^+ , we use the reference translation of input $\mathbf{f}_i$. For the contrast set d^-, we sample another sentence from the target side of the training data. We train on 500-best lists for 20 epochs. We conduct this experiment on the IWSLT evaluation data, using IWSLT tst2010 for training and tst2013 for evaluation. The model is trained on out-of-domain data in the same way as the model described in 4.1. Figure 1 shows learning curves for L_{ramp_1} which uses cost and L_{ramp_2} which uses $\mathrm{cost\_diff}$. We found $\mathrm{cost\_diff}$ to perform much better than cost on both train and heldout data. Why does cost do so much worse? Remember, that in this scenario we select two *hope*-derivations. What is more, we only select them from a small k-best list (500 translations). With the $\mathrm{cost\_diff}$ function we are required to choose hypotheses that *distinguish* most between d^+ and d^-. This will select a h^- that is far away from the reference (similar to a *fear* derivation). While with cost, there is no guarantee that h^- will differ from h^+.

3 Data preparation and extraction

3.1 Initial Wikipedia data set

Wikipedia is internally structured by cross-lingual links and inter-article links. We use the German-English WikiCLIR collection by Schamoni et al. (2014), along with their definition of cross-lingual relevance levels: A target language document has relevance level 3 if it is the cross-lingual mate of an input document. It is assigned relevance level 2, if there is a bidirectional link relation between the cross-lingual mate and the document. WikiCLIR contains a total of 225,294 mate relations with 1 average German mate per English document, and over 1.7 million bidirectional link relations, with on average 8.5 links per English document. We use the link information provided by WikiCLIR, but we work with the full Wikipedia documents rather than WikiCLIR's abbreviated queries.

3.2 Automatic sentence alignment

The cross-lingual mate relation in Wikipedia is a strong indicator for parallelism. However, Wikipedia entries in different languages are not necessarily translations of each other, but can be edited independently. In order to find parallel sentences, we use an automated extraction method. We do this for three purposes:

1. To identify nearly parallel document pairs for the construction of a clean in-domain evaluation set without having to rely on manual translation.

2. To examine whether bidirectional links provide a strong enough signal for extracting pseudo-parallel training data.

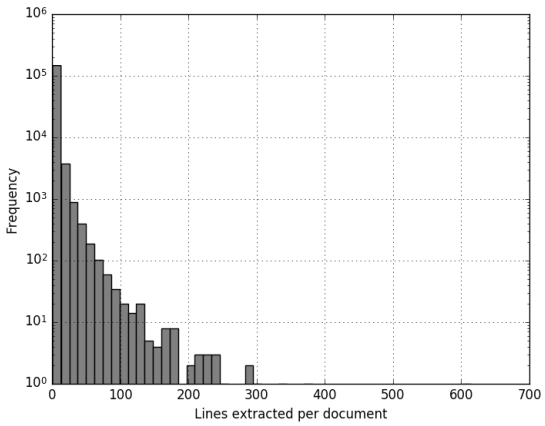

Figure 2: Number of documents (y-axis, on log-scale) from which n lines were extracted (x-axis).

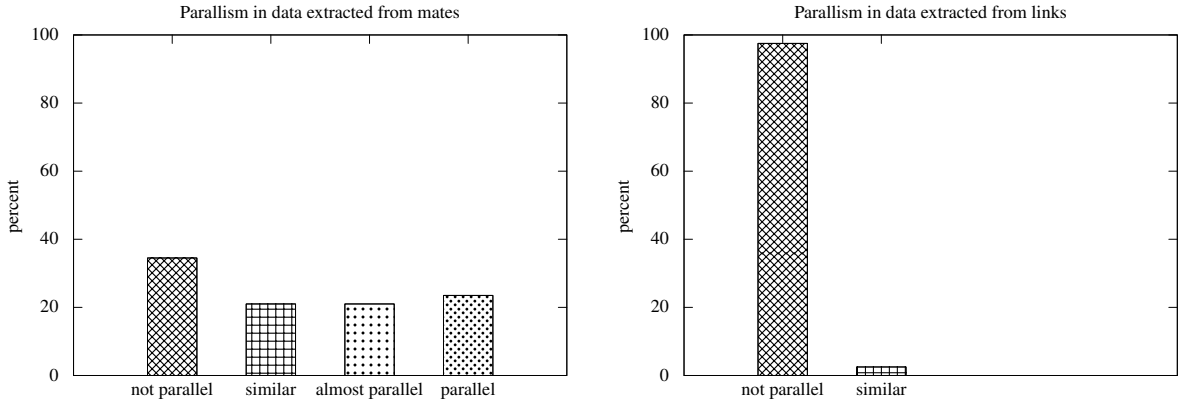

Figure 3: Sentence aligner precision for mates and links.

3. To compare our method to automatic parallel data extraction based on relevance annotation.

We use the modified `yalign` method described by Wołk and Marasek (2015) for pseudo-parallel data extraction.We adapt their software to handle the WikiCLIR format. `yalign` requires a bilingual dictionary with translation probabilities. Following Wołk and Marasek (2015), we use a lexical translation table created from the TED parallel training data[2] as our bilingual dictionary. We filter the dictionary for punctuation and numerals and discard all entries whose lexical translation probability is smaller than 0.3.

Figure 2 shows the frequency histogram of the number of extracted lines per document pair for document pairs with a mate relation. For most document pairs, only a single sentence pair was extracted. However, there were a few document pairs that yielded several hundred sentence pairs. In total, 533,516 sentence pairs were extracted.

Figure 3 shows our evaluation of `yalign`'s precision for the mate and link relations. We manually evaluated a sample of 200 automatically aligned sentence pairs. The sentence pairs were annotated using four categories: "fully parallel", "almost parallel" – this category contains sentence pairs that have parallel segments, with other segments missing from the aligned part, "similar" – for sentence pairs that have similar content or wording but differ factually –, and "non parallel". While 65.5% of sentence pairs from the mate relation were similar or parallel, the link relation yielded only 2.6% sentence pairs that were at least similar. We conclude that the bidirectional link relation is too weak to extract useful pseudo-parallel data.

[2] `https://wit3.fbk.eu/`

Set	Length	# parallel	Title
set1	323	285	Polish culture during World War II
	710	677	Black-figure pottery
	457	375	Ulm Hauptbahnhof
	587	375	Characters of Carnivàle
Total		1712	
set2	360	268	J-pop
	501	388	Schüttorf
	549	438	Military history of Australia during World War II
	676	432	Arab citizens of Israel
Total		1526	

Table 1: Wikipedia development and test documents.

3.3 Evaluation data construction

To construct our in-domain evaluation data, we sorted all automatically aligned documents by the number of aligned sentences up to a limit of 10,000 sentences. We then selected eight documents for manual alignment, discarding other document pairs which appeared to have been machine-translated, only contained few parallel sentences, or consisted of lists of proper names. During manual alignment, we also fixed sentence splitting errors and removed image captions and references. We split the documents into two groups of four, making sure to keep the sets diverse. Table 1 shows the two sets of extracted documents. They are topically diverse, similar in length, and contain a considerable percentage of parallel sentences.

4 Experiments

4.1 Out-of-domain translation system

Our baseline English-German translation system is trained on 2.1 million sentence pairs (61/59 million English/German tokens) from the Europarl v7[3] corpus (1.78 million sentence pairs), the News Commentary v10[4] corpus (200K sentence pairs) and the MultiUN v1[5] corpus (150K sentence pairs). Word alignments are computed using MGIZA++[6], alignments are symmetrized using the `grow-diag-final-end` heuristic. A 4-gram count-based language model is estimated from the target side of the training data using `lmplz` (Heafield et al., 2013). All experiments use the hierarchical phrase-based decoder `cdec` (Dyer et al., 2010). Hierarchical phrase rules are extracted using `cdec`'s implementation of the suffix array extractor by Lopez (2007) with default settings. Our baselines use 21 decoder features (7 translation model features, 2 language model features, 7 pass through features, 3 arity penalty features, word penalty and glue rule count features), which are implemented in `cdec`. Feature weights are optimized on the WMT Newstest 2014 data set (3003 sentence pairs) using the pairwise ranking optimizer `dtrain`[7]. We run `dtrain` for 15 epochs with the hyperparameters k-best size=100, loss-margin=1, and a learning rate of $1e^{-5}$. The final weights are averaged over all epochs. Performance of our baseline system (*baseline 1*) is given in the first row of Table 2.

4.2 Translation model and language model adaptation

For translation model (TM) adaptation we add the automatically extracted pseudo-parallel Wikipedia data (see Section 3) to our baseline training data and re-train the translation model. For language model (LM) adaptation, we sample 500,000 sentences from the German Wikipedia data, which we add to the out-of-domain language model data to re-build a combined 4-gram language model. Both language model and translation model adaptation boosted performance. Rows 1 and 3 in Table 3 show BLEU

[3] `www.statmt.org/europarl/`, see (Koehn, 2005)
[4] `www.statmt.org/wmt15/training-parallel-nc-v10.tgz`
[5] `www.euromatrixplus.net/multi-un/`, see (Eisele and Chen, 2010)
[6] www.cs.cmu.edu/ qing/giza/
[7] `https://github.com/pks/cdec-dtrain`

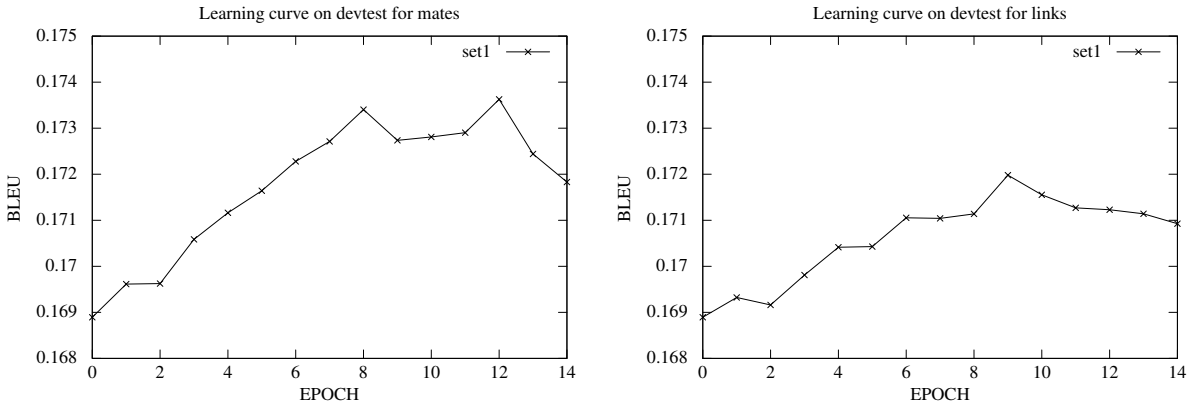

Figure 4: Performance on heldout *set1* for mates and links.

Experiment	%BLEU *set2*
baseline 1 (out-of-domain)	12.46
kbest-train (mates, cost_diff)	12.57 (+0.11)
hypergraph-train (mates, cost_diff)	13.05* (+0.59)
hypergraph-train (mates, cost)	12.81* (+0.34)
hypergraph-train (mates+links, cost_diff)	12.85* (+0.38)
hypergraph-train (links, cost_diff, random sampling)	12.67 (+0.21)
hypergraph-train (links, cost_diff, weighted sampling)	12.77* (+0.31)

Table 2: Results for training on Wikipedia with out-of-domain model. * indicates a significant difference to the baseline at a significance level of 0.05.

scores for an LM-adapted model (*baseline 2*) and a model with both LM and TM adaptation (*baseline 3*). The good performance of TM adaptation leads to the conclusion that if there is a strong signal for potential parallelism like in the Wikipedia data, automatic pseudo-parallel data extraction works well.

4.3 Learning from Wikipedia mates and links

We train our method on 10,000 input sentences sampled from the English WikiCLIR documents. Each input sentence is annotated with a document identifier in order to sample positive and negative examples. The relevant document collection D^+ includes all German documents which are linked to an English document by a mate or bidirectional link relation. For the contrast documents D^- we use the News Commentary corpus, split into documents. In a pre-processing step, we extract n-grams up to order 3 from each document, which we need to calculate n-gram precision. We also experimented with a larger training set of 200,000 input sentences but found no significant improvement. All experiments use the same 21 features as the baseline, keeping those weights fixed, but train sparse lexicalized features (rule identifiers, rule source and target bigrams and lexical alignment features described in Simianer et al. (2012)) in parallel on 10 shards, followed by an ℓ_1/ℓ_2 feature selection step which keeps at most 100,000 features. We use a constant learning rate of $\eta = 1e^{-4}$ and regularization strength $C = 1$. Experiments were run for up to 20 epochs and performance on the heldout *set1* was used as an early stopping criterion.

Table 2 reports BLEU scores on *set2*, when our model is trained on an unadapted baseline model (*baseline 1*). Significance tests were conducted by `multeval` (Clark et al., 2011). While training on k-best lists produced a small incremental gain, training on hypergraphs improved up to 0.6 BLEU over the baseline. Both cost and cost_diff produced an improvement over the baseline, however, cost_diff performed slightly better. As expected, using the strongest signal, the cross-lingual mate relation, worked best. When only the link relation was used, only the experiment with the weighted sampling strategy produced a significant improvement. Figure 4 shows learning curves over epochs on heldout *set1* for

Experiment	%BLEU *set2*
baseline 2 (LM adaptation)	13.62
hypergraph-train (mates, cost_diff)	13.93* (+0.31)
baseline 3 (LM and TM adaptation)	14.96
hypergraph-train (mates, cost_diff)	15.17* (+0.21)

Table 3: Results for training on Wikipedia with adapted model. * indicates a significant difference to the baseline at a significance level of 0.05.

training on mates and links (both with random sampling).

Table 3 shows results for training on mates with an adapted baseline model. In both experiments, there was a small, yet significant, improvement over the adapted model, showing that additional information can be learned from the relevance signal.

Examples. To give a better impression what is learned by our method, Table 4 contains some example translations from *baseline 1* and the best adapted model from Table 2. Spans in which our model performed better are marked in **boldface**. *Example 1* shows that our model fixed word order mistakes made by the baseline, such as "Apartheid zionistischen", which is fixed to "zionistischen Apartheid". The same is true for the proper name and attribution "Thomas Michael Hamerlik (CDU)" in *Example 2*. Both examples suggest that by training on Wikipedia documents, which include frequent parentheses, quotations and named entities, our model becomes better at handling these types of phrases. *Example 3* is interesting, because in this case the baseline produced an idiomatic, rather informal translation for "postponed indefinitely" ("auf den Sankt - Nimmerleins - Tag verschoben") which would be correct in a spoken language context but strange to use in a Wikipedia article, while our model produced the correct translation ("auf unbestimmte Zeit verschoben").

5 Conclusion and future work

In this paper we have presented a new objective for learning translation model parameters from graded and negative relevance signals. Using Wikipedia translation as an example, we were able to achieve significant improvements over an unadapted baseline. As expected, a stronger relevance signal produced larger gains, but we were able to produce small, but significant, improvements, even when learning only from indirect links. We compare our method to baselines that use monolingual data to adapt the language model or rely on strong parallelism signals to adapt the translation model. Our approach was able to yield a small gain even when combined with these strong baselines.

It is worth mentioning that our approach is not restricted to Wikipedia data, but could be applied to other large multilingual collections where cross-lingual relevance information can be extracted. For example, cross-lingual mates could be extracted for multilingual patent corpora through patent family relations (i.e. versions of the same patent submitted to different patent organizations). In addition, weaker links are given by the international patent classification system, or by citations between patents. Another application scenario could be social media data which use the same hashtags across languages. If no explicit signals are available, or if they are not strong enough, one could also use unsupervised document similarity metrics or cross-language information retrieval techniques to detect relevant documents in a target language collection and use these documents as positive examples. We plan to explore these directions in the future.

Since our general learning setup and objective is agnostic about the type of translation system we also plan to apply it to neural machine translation.

Acknowledgements

This research was supported in part by DFG grant RI-2221/1-2 "Weakly Supervised Learning of Cross-Lingual Systems".

Example 1	
Source	political demands include " **the return of all Palestinian refugees to their homes and lands** , [an] end [to] the Israeli occupation and **Zionist apartheid** and the establishment [of] a democratic secular state in Palestine as the ultimate solution to the Arab - Zionist conflict . "
Baseline 1	politische Forderungen : " **alle palästinensischen Flüchtlingen die Rückkehr an ihre Heimstätten und Land beenden** , [an] [. . .] der israelischen Besatzung und **Apartheid zionistischen** [der] sowie die Einrichtung einer demokratischen säkularen Staat in Palästina als die ultimative Lösung für das arabisch - zionistischen Konflikt . "
Hypergraph-train	politische Forderungen aufzunehmen " **die Rückkehr aller palästinensischen Flüchtlinge in ihre Heimat und zu ihren Ländereien** , [an] Ende [. . .] der israelischen Besatzung und **zionistischen Apartheid** und die Einrichtung [der] einen demokratischen säkularen Staat in Palästina als die ultimative Lösung des arabisch - zionistischen Konflikt . "
Reference	politische Forderungen von Abnaa el-Balad sind u. a. " ... **die Rückkehr aller palästinensischen Flüchtlinge in ihre Heimat und auf ihr Land** , [ein] Ende [der] israelischen Besatzung und **zionistischen Apartheid** und die Gründung eines demokratischen säkularen Staates in Palästina als endgültige Lösung des arabisch - zionistischen Konflikts .
Example 2	
Source	the current mayor is **Thomas Michael Hamerlik (CDU)** with two deputies :
Baseline 1	der derzeitige Bürgermeister **Michael Hamerlik Thomas** ist mit zwei Stellvertreter **(CDU)** :
Hypergraph-train	der derzeitige Bürgermeister ist **Thomas Michael Hamerlik (CDU)** mit zwei Abgeordneten :
Reference	Bürgermeister ist zurzeit **Thomas Michael Hamerlik (CDU)** mit zwei Stellvertretern :
Example 3	
Source	this plan was frustrated by the Japanese defeat in the Battle of the Coral Sea and was **postponed indefinitely** after the Battle of Midway .
Baseline 1	dieser Plan wurde von den Japanern frustriert Niederlage im Kampf der Coral See und nach der Schlacht von Midway **auf den Sankt - Nimmerleins - Tag verschoben** wurde .
Hypergraph-train	dieser Plan wurde frustriert durch die japanische Niederlage im Kampf der Coral Meer und nach der Schlacht von Midway **auf unbestimmte Zeit verschoben** wurde .
Reference	der japanische Plan erlitt mit der Niederlage in der Schlacht im Korallenmeer einen ersten Rückschlag und wurde nach der Niederlage in der Schlacht um Midway **auf unbestimmte Zeit verschoben** .

Table 4: Translation examples from the test set, comparing the unadapted baseline to adaptation with our method.

References

David Chiang. 2012. Hope and fear for discriminative training of statistical translation models. *Journal of Machine Learning Research*, 13(Apr):1159–1187.

Jonathan Clark, Chris Dyer, Alon Lavie, and Noah Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Lingustics*, ACL '11, Portland, Oregon, USA.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, Uppsala, Sweden.

Pavlos S. Efraimidis and Paul G. Spirakis. 2006. Weighted random sampling with a reservoir. *Information Processing Letters*, 97(5):181 – 185.

Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In *LREC*.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for smt. In *Proceedings of the Second Workshop on Statistical Machine Translation*, WMT '07, Prague, Czech Republic.

Kevin Gimpel and Noah A. Smith. 2012. Structured ramp loss minimization for machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, Montreal, Canada.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL '13, Sofia, Bulgaria.

Joseph Keshet and David A McAllester. 2011. Generalization bounds and consistency for latent structural probit and ramp loss. In *Advances in Neural Information Processing Systems*, pages 2205–2212.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, WMT '07, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Adam Lopez. 2007. Hierarchical phrase-based translation with suffix arrays. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado, June. Association for Computational Linguistics.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Shigehiko Schamoni, Felix Hieber, Artem Sokolov, and Stefan Riezler. 2014. Learning translational and knowledge-based similarities from relevance rankings for cross-language retrieval. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL '14, Baltimore, MD, USA.

Patrick Simianer, Stefan Riezler, and Chris Dyer. 2012. Joint feature selection in distributed stochastic learning for large-scale discriminative training in smt. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ACL '12, Jeju Island, Korea.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL '13, Sofia, Bulgaria.

Krzysztof Wołk and Krzysztof Marasek. 2015. Unsupervised comparable corpora preparation and exploration for bi-lingual translation equivalents. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, IWSLT '15, Da Nang, Vietnam.