Enhancing Morphological Alignment for Translating Highly Inflected Languages *

Minh-Thang Luong
School of Computing
National University of Singapore
luongmin@comp.nus.edu.sg

Min-Yen Kan
School of Computing
National University of Singapore
kanmy@comp.nus.edu.sg

Abstract

We propose an unsupervised approach utilizing only raw corpora to enhance morphological alignment involving highly inflected languages. Our method focuses on *closed-class morphemes*, modeling their influence on nearby words. Our language-independent model recovers important links missing in the IBM Model 4 alignment and demonstrates improved end-to-end translations for English-Finnish and English-Hungarian.

1 Introduction

Modern statistical machine translation (SMT) systems, regardless of whether they are word-, phrase- or syntax-based, typically use the word as the atomic unit of translation. While this approach works when translating between languages with limited morphology such as English and French, it has been found inadequate for morphologically-rich languages like Arabic, Czech and Finnish (Lee, 2004; Goldwater and McClosky, 2005; Yang and Kirchhoff, 2006). As a result, a line of SMT research has worked to incorporate morphological analysis to gain access to information encoded within individual words.

In a typical MT process, word aligned data is fed as training data to create a translation model. In cases where a highly inflected language is involved, the current word-based alignment approaches produce low-quality alignment, as the statistical correspondences between source and target words are diffused over many morphological forms. This problem has a direct impact on end translation quality.

Our work addresses this shortcoming by proposing a morphologically sensitive approach to word alignment for language pairs involving a highly inflected language. In particular, our method focuses on a set of *closed-class morphemes* (CCMs), modeling their influence on nearby words. With the model, we correct erroneous alignments in the initial IBM Model 4 runs and add new alignments, which results in improved translation quality.

After reviewing related work, we give a case study for morpheme alignment in Section 3. Section 4 presents our four-step approach to construct and incorporate our CCM alignment model into the grow-diag process. Section 5 describes experiments, while Section 6 analyzes the system merits. We conclude with suggestions for future work.

2 Related Work

MT alignment has been an active research area. One can categorize previous approaches into those that use language-specific syntactic information and those that do not. Syntactic parse trees have been used to enhance alignment (Zhang and Gildea, 2005; Cherry and Lin, 2007; DeNero and Klein, 2007; Zhang et al., 2008; Haghighi et al., 2009). With syntactic knowledge, modeling long distance reordering is possible as the search space is confined to plausible syntactic variants. However, they generally require language-specific tools and annotated data, making such approaches infeasible for many languages. Works that follow non-syntactic approaches, such as (Matusov et al.,

*This work was supported by a National Research Foundation grant "Interactive Media Search" (grant # R-252-000-325-279)

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 743–751,
Beijing, August 2010

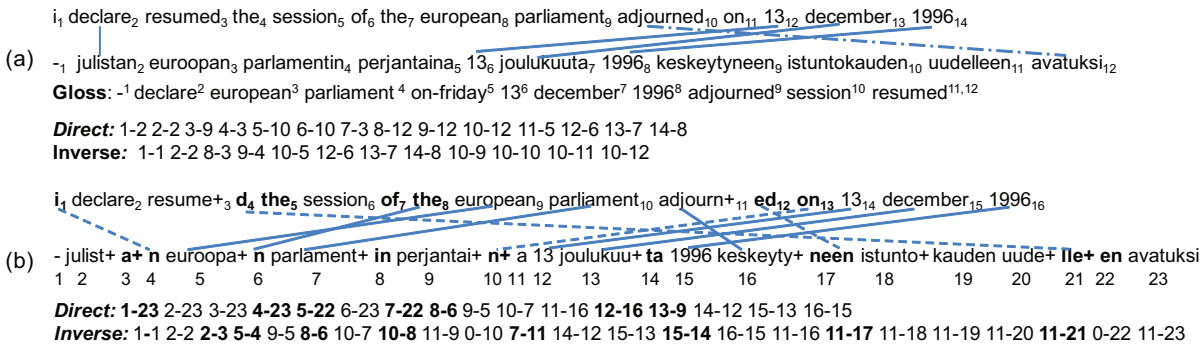

Figure 1: **Sample English-Finnish IBM Model 4 alignments**: (a) word-level and (b) morpheme-level. Solid lines indicate intersection alignments, while the exhaustive asymmetric alignments are listed below. In (a), translation glosses for Finnish are given; the dash-dot line is the incorrect alignment. In (b), bolded texts are closed-class morphemes (CCM), while bolded indices indicate alignments involving CCMs. The dotted lines are correct CCM alignments not found by IBM Model 4.

2004; Liang et al., 2006; Ganchev et al., 2008), which aim to achieve symmetric word alignment during training, though good in many cases, are not designed to tackle highly inflected languages.

Our work differs from these by taking a middle road. Instead of modifying the alignment algorithm directly, we preprocess asymmetric alignments to improve the input to the symmetrizing process later. Also, our approach does not make use of specific language resources, relying only on unsupervised morphological analysis.

3 A Case for Morpheme Alignment

The notion that morpheme based alignment would be useful in highly inflected languages is intuitive. Morphological inflections might indicate tense, gender or number that manifest as separate words in largely uninflected languages. Capturing these subword alignments can yield better word alignments that otherwise would be missed.

Let us make this idea concrete with a case study of the benefits of morpheme based alignment. We show the intersecting alignments of an actual English (source) $\rightarrow$ Finnish (target) sentence pair in Figure 1, where (a) word-level and (b) morpheme-level alignments are shown. The morpheme-level alignment is produced by automatically segmenting words into morphemes and running IBM Model 4 on the resulting token stream.

Intersection links (*i.e.*, common to both direct and inverse alignments) play an important role in creating the final alignment (Och and Ney, 2004). While there are several heuristics used in the symmetrizing process, the *grow-diag(onal)* process is common and prevalent in many SMT systems, such as Moses (Koehn et al., 2007). In the grow-diag process, intersection links are used as seeds to find other new alignments within their neighborhood. The process continues iteratively, until no further links can be added.

In our example, the morpheme-level intersection alignment is better as it has no misalignments and adds new alignments. However it misses some key links. In particular, the alignments of closed-class morphemes (CCMs; later formally defined) as indicated by the dotted lines in (b) are overlooked in the IBM Model 4 alignment. This difficulty in aligning CCMs is due to:

1. Occurrences of *garbage-collector* words (Moore, 2004) that attract CCMs to align to them. Examples of such links in (b) are 1–23 or 11–21 with the occurrences of rare words `adjourn+`$_{11}$ and `avatuksi`$_{23}$. We further characterize such errors in Section 6.1.

2. Ambiguity among CCMs of the same surface that causes incorrect matchings. In (b), we observe multiple occurrence of `the` and `n` on the source and target sides respectively. While the link 8–6 is correct, 5–4 is not as $\mathtt{i}_1$ should be aligned to $\mathtt{n}_4$ instead. To resolve such ambiguity, context information should be considered as detailed in Section 4.3.

The fact that rare words and multiple affixes often occur in highly inflected languages exacerbates this problem, motivating our focus on improving CCM alignment. Furthermore, having access to the correct CCM alignments as illustrated

in Figure 1 guides the grow-diag process in finding the remaining correct alignments. For example, the addition of CCM links i$_1$–n$_4$ and d$_4$–lle$_{21}$ helps to identify declare$_2$–julist$_2$ and resume$_3$–avatuksi$_{23}$ as admissible alignments, which would otherwise be missed.

4 Methodology

Our idea is to enrich the standard IBM Model 4 alignment by modeling closed-class morphemes (CCMs) more carefully using global statistics and context. We realize our idea by proposing a four-step method. First, we take the input parallel corpus and convert it into morphemes before training the IBM Model 4 morpheme alignment. Second, from the morpheme alignment, we induce automatically bilingual CCM pairs. The core of our approach is in the third and fourth steps. In Step 3, we construct a *CCM alignment model*, and apply it on the segmented input corpus to obtain an automatic CCM alignment. Finally, in Step 4, we incorporate the CCM alignment into the symmetrizing process via our *modified grow-diag process*.

4.1 Step 1: Morphological Analysis

The first step presupposes morphologically segmented input to compute the IBM Model 4 morpheme alignment. Following Virpioja et al. (2007), we use *Morfessor*, an unsupervised analyzer which learns morphological segmentation from raw tokenized text (Creutz and Lagus, 2007).

The tool segments input words into labeled morphemes: PRE (prefix), STM (stem), and SUF (suffix). Multiple affixes can be proposed for each word; word compounding is allowed as well, e.g., uncarefully is analyzed as un/PRE+ care/STM+ ful/SUF+ ly/SUF. We append a "+" sign to each nonfinal tag to distinguish word-internal morphemes from word-final ones, e.g., "x/STM" and "x/STM+" are considered different tokens. The "+" annotation enables the restoration of the original words, a key point to enforce word boundary constraints in our work later.

4.2 Step 2: Bilingual CCM Pairs

We observe that low and highly inflected languages, while intrinsically different, share more in common at the morpheme level. The many-to-one relationships among words on both sides is often captured better by one-to-one correspondences among morphemes. We wish to model such bilingual correspondence in terms of closed-class morphemes (CCM), similar to Nguyen and Vogel (2008)'s work that removes nonaligned affixes during the alignment process. Let us now formally define CCM and an associative measure to gauge such correspondence.

en	fi	en	fi	en	fi
the$_1$	-n$_1^\dagger$	in$_6$	-ssa$_{15}^\ddagger$	me$_{166}$	-ni$_{60}^\ddagger$
-s$_2$	-t$_9^\ddagger$	is$_7$	on$_2^\ddagger$	me$_{166}$	minun$_{282}^\dagger$
to$_3$	-ä$_6$	that$_8$	että$_7^\ddagger$	why$_{168}$	siksi$_{187}^\ddagger$
to$_3$	maan$_{91}$	that$_8$	ettei$_{283}^\ddagger$	view$_{172}$	mieltä$_{162}^\dagger$
of$_4$	-a$_4$	we$_{10}$	-mme$_{10}^\ddagger$	still$_{181}$	vielä$_{108}^\ddagger$
of$_4$	-en$_5^\dagger$	we$_{10}$	meidän$_{52}^\dagger$	where$_{183}$	jossa$_{209}^\ddagger$
of$_4$	-sta$_{19}^\dagger$	we$_{10}$	me$_{113}^\ddagger$	same$_{186}$	samaa$_{334}^\ddagger$
and$_5$	ja$_3^\ddagger$	we$_{10}$	emme$_{123}$	he$_{187}$	hän$_{184}^\ddagger$
and$_5$	sekä$_{122}^\ddagger$	we$_{10}$	meillä$_{231}^\dagger$	good$_{189}$	hyvä$_{321}^\ddagger$
and$_5$	eikä$_{203}$	...	...	over-$_{408}$	yli-$_{391}^\ddagger$

Table 1: **English(en)-Finnish(fi) Bilingual CCM pairs** (N=128). Shown are the top 19 and last 10 of 168 bilingual CCM pairs extracted. Subscript i indicates the i^{th} most frequent morpheme in each language. ‡ marks exact correspondence linguistically, whereas † suggests rough correspondence w.r.t http://en.wiktionary.org/wiki/.

Definition 1. *Closed-class Morphemes (CCM)* are a *fixed set of stems and affixes* that exhibit grammatical functions just like closed-class words. In highly inflected languages, we observe that grammatical meanings may be encoded in morphological stems and affixes, rather than separate words. While we cannot formally identify valid CCMs in a language-independent way (as by definition they manifest language-dependent grammatical functions), we can devise a good approximation. Following Setiawan et al. (2007), we induce the set of CCMs for a language as the top N frequent stems together with all affixes[1].

Definition 2. *Bilingual Normalized PMI (biPMI)* is the averaged normalized PMI computed on the asymmetric morpheme alignments. Here, normalized PMI (Bouma, 2009), known to be less biased towards low-frequency data, is defined as: $nPMI(x,y) = \ln \frac{p(x,y)}{p(x)p(y)} / -\ln p(x,y)$, where $p(x)$, $p(y)$, and $p(x,y)$ follow definitions in the standard PMI formula. In our case, we only

[1]Note that we employ length and vowel sequence heuristics to filter out corpus-specific morphemes.

compute the scores for x, y being morphemes frequently aligned in both asymmetric alignments.

Given these definitions, we now consider a pair of source and target CCMs related and termed a **bilingual CCM pair** (CCM pair, for short) if they exhibit positive correlation in their occurrences (*i.e.*, positive nPMI[2] and frequent cooccurrences).

We should note that relying on a hard threshold of N as in (Setiawan et al., 2007) is brittle as the CCM set varies in sizes across languages. Our method is superior in the use of N as a starting point only; the bilingual correspondence of the two languages will ascertain the final CCM sets.

Take for example the *en* and *fi* CCM sets with 154 and 214 morphemes initially (each consisting of N=128 stems). As morphemes not having their counterparts in the other language are spurious, we remove them by retaining only those in the CCM pairs. This effectively reduces the respective sizes to 91 and 114. At the same time, these final CCMs cover a much larger range of top frequent morphemes than N, up to 408 *en* and 391 *fi* morphemes, as evidenced in Table 1.

4.3 Step 3: The CCM Alignment Model

The goal of this model is to predict when appearances of a CCM pair should be deemed as linking.

With an identified set of CCM pairs, we know when source and target morphemes correspond. However, in a sentence pair there can be many instances of both the source and target morphemes. In our example, the `the`–`n` pair corresponds to definite nouns; there are two `the` and three `-n` instances, yielding 2×3=6 possible links.

Deciding which instances are aligned is a decision problem. To solve this, we inspect the IBM Model 4 morpheme alignment to construct a CCM alignment model. The CCM model labels whether an instance of a CCM pair is deemed semantically related (linked). We cast the modeling problem as supervised learning, where we choose a maximum entropy (ME) formulation (Berger et al., 1996).

We first discuss sample selection from the IBM Model 4 morpheme alignment, and then give details on the features extracted. The processes described below are done per sentence pair with f_1^m, e_1^n and U denoting the source, target sentences and the union alignments, respectively.

Class labels. We base this on the initial IBM Model 4 alignment to label each CCM pair instance as a positive or negative example: *Positive examples* are simply CCM pairs in U. To be precise, links j–i in U are positive examples if f_j–e_i is a CCM pair. To find *negative examples*, we inventory other potential links that share the same lexical items with a positive one. That is, a link j'–i' not in U is a negative example, if a positive link j–i such that $f_j = f'_j$ and $e_i = e'_i$ exists.

We stress that our collection of positive examples contains high-precision but low-recall IBM Model 4 links, which connect the reliable CCM pairs identified before. The model then generalizes from these samples to detect incorrect CCM links and to recover the correct ones, enhancing recall. We later detail this process in §4.4.

Feature Set. Given a CCM pair instance, we construct three feature types: lexical, monolingual, and bilingual (See Table 2). These features capture the global statistics and contexts of CCM pairs to decide if they are true alignment links.

• **Lexical features** reflect the tendency of the CCM pair being aligned to themselves. We use *biPMI*, which aggregates the global alignment statistics, to determine how likely source and target CCMs are associated with each other.

• **Monolingual context features** measure the association among tokens of the same language, capturing what other stems and affixes co-occur with the source/target CCM:

1. within the same word (*intra*). The aim is to disambiguate affixes as necessary in highly inflected languages where same stems could generate different roles or meanings.
2. outside the CCM's word boundary (*inter*). This potentially capture cues such as tense, or number agreement. For example, in English, the 3sg agreement marker on verbs `-s` often co-occurs with nearby pronouns e.g., `he`, `she`, `it`; whereas the same marker on nouns (`-s`), often appears with plural determiners e.g., `these`, `those`, `many`.

To accomplish this, we compute two monolingual nPMI scores in the same spirit as biPMI, but using the morphologically segmented input from

[2] *nPMI* has a bounded range of $[-1, 1]$ with values 1 and 0 indicating perfect positive and no correlation, respectively.

Feature Description	Examples
Lexical — *biPMI*: None $[-1, 0]$, Low $(0, 1/3]$, Medium $(1/3, 2/3]$, High $(2/3, 1]$	pmi_{d-lle}=Low
Monolingual Context — Capture morpheme cooccurrence with the src/tgt CCM	
Intra – **W**ithin the same word	srcW_{d-lle}=resume, tgtW_{d-lle}=en, tgtW_{d-lle}=uude
Inter – To the **L**eft & **R**ight, in different words	srcL_{d-lle}=i, srcR_{d-lle}=the, tgtR_{d-lle}=avatuksi
Bilingual context — Capture neighbor links' cooccurrence with the CCM pair link	
bi0 – Most descriptive, capturing in terms of surface forms only → maybe sparse	bi0_{d-lle}=resume–avatuksi
bi1 – Generalizes morphemes into relative locations (**L**eft, **W**ithin, **R**ight)	bi1_{d-lle}=W–avatuksi, bi1_{d-lle}=resume–R
bi2 – Most general, coupling token types (**C**lose, **O**pen) /w relative positions	bi2_{d-lle}=O–WR

Table 2: **Maximum entropy feature set.** Shown are feature types, descriptions and examples. Most examples are given for the alignment $\texttt{d}_4\texttt{–lle+}_{21}$ of the same running example in §3. Note that we only partially list the bilingual context features.

each language separately. Two morphemes are "linked" if within a context window of w_c words.

• **Bilingual context features** model cross-lingual reordering, capturing the relationships between the CCM pair link and its neighbor[3] links. Consider a simple translation between an English phrase of the form `we` ⟨`verb`⟩ and the Finnish one ⟨`verb`⟩ `–mme`, where `–mme` is the 1pl verb marker. We aim to capture movements such as "the open-class morphemes on the right of `we` and on the left of `–mme` are often aligned". These will function as evidence for the ME learner to align the CCM pair (`we`, `–mme`). We encode the bilingual context at three different granularities, from most specific to most general ones (cf Table 2).

4.4 Step 4: Incorporate CCM Alignment

At test time, we apply the trained CCM alignment model to all CCM pairs occurring in each sentence pair to find CCM links. On our running example in Figure 1, the CCM classifier tests 17 CCM pairs, identifying 6 positive CCM links of which 4 are true positives (dotted lines in (b)).

Though mostly correct, we note that some of the predicted links conflict: ($\texttt{d}_4\texttt{–lle}_{21}$, $\texttt{ed}_{12}\texttt{–neen}_{17}$ and $\texttt{ed}_{12}\texttt{–lle}_{21}$) share alignment endpoints. Such sharing in CCM alignments is rare and we believe should be disallowed. This motivates us to resolve all CCM link conflicts before incorporating them into the symmetrizing process.

Resolving link conflicts. As CCM pairs are classified independently, they possess classification probabilities which we use as evidence to resolve the conflicts. In our example, the classification probabilities for ($\texttt{d}_4\texttt{–lle}_{21}$, $\texttt{ed}_{12}\texttt{–neen}_{17}$, $\texttt{ed}_{12}\texttt{–lle}_{21}$) are $(0.99, 0.93, 0.79)$ respectively.

We use a simple, "best-first" greedy approach to determine which links are kept and which are dropped to satisfy our assumption. In our case, we pick the most confident link, $\texttt{d}_4\texttt{–lle}_{21}$ with probability 0.99. This precludes the incorrect link, $\texttt{ed}_{12}\texttt{–lle}_{21}$, but admits the other correct one $\texttt{ed}_{12}\texttt{–neen}_{17}$, probability 0.93. As a result, this resolution successfully removes the incorrect link.

Modifying grow-diag. We incorporate the set of conflict-resolved CCM links into the grow-diag process. This step modifies the input alignments as well as the growing process. U and I denote the IBM Model 4 union and intersection alignments.

In our view, the resolved CCM links can serve as a quality mark to "upgrade" links before input into the grow-diag process. We upgrade resolved CCM links: (a) those $\in U \rightarrow$ part of I, treating them as alignment seeds; (b) those $\notin U \rightarrow$ part of U, using them for exploration and growing. To reduce spurious alignments, we discarded links in U that conflict with the resolved CCM links.

In the usual grow-diag, links immediately adjacent to a seed link l are considered candidates to be appended into the alignment seeds. While suitable for word-based alignment, we believe it is too small a context when the input are morphemes.

For morpheme alignment, the candidate context makes more sense in terms of word units. We thus *enforce word boundaries* in our modified grow-diag. We derive word boundaries for end points in l using the morphological tags and the "+" word-end marker mentioned in §4.1. Using such boundaries, we can then extend the grow-diag to consider candidate links within a neighborhood of w_g words; hence, enhancing the candidate coverage.

5 Experiments

We use English-Finnish and English-Hungarian data from past shared tasks (WPT05 and WMT09)

[3] Within a context window of w_c words as in monolingual.

to validate our approach. Both Finnish and Hungarian are highly inflected languages, with numerous verbal and nominal cases, exhibiting agreement. Dataset statistics are given in Table 3.

	en-fi	#	en-hu	#
Train	Europarl-v1	714K	Europarl-v4	1,510K
LM	Europarl-v1-fi	714K	News-hu	4,209K
Dev	wpt05-dev	2000	news-dev2009	2051
Test	wpt05-test	2000	news-test2009	3027

Table 3: **Dataset Statistics:** the numbers of parallel sentences for training, LM training, development and test sets.

We use the Moses SMT framework for our work, creating both our CCM-based systems and the baselines. In all systems built, we obtain the IBM Model 4 alignment via GIZA++ (Och and Ney, 2003). Results are reported using case-insensitive BLEU (Papineni et al., 2001).

Baselines. We build two SMT baselines:

w-system: This is a standard phrase-based SMT, which operates at the word level. The system extracts phrases of maximum length 7 words, and uses a 4-gram word-based LM.

w_m-system: This baseline works at the word level just like the w-system, but differs at the alignment stage. Specifically, input to the IBM Model 4 training is the morpheme-level corpus, segmented by *Morfessor* and augmented with "+" to provide word-boundary information (§4.1). Using such information, we constrain the alignment symmetrization to extract phrase pairs of 7 words or less in length. The morpheme-based phrase table is then mapped back to word forms. The process continues identically as in the w-system.

CCM-based systems. Our CCM-based systems are similar in spirit to the w_m system: train at the morpheme, but decode at the word level. We further enhance the w_m-system at the alignment stage. First, we train our CCM model based on the initial IBM Model 4 morpheme alignment, and apply it to the morpheme corpus to obtain CCM alignment, which are input to our modified grow-diag process. The CCM approach defines the setting of three parameters: $\langle N, w_c, w_g \rangle$ (Section 4). Due to our resource constraints, we set N=128, similar to (Setiawan et al., 2007), and w_c=1 experimentally. We only focus on the choice of w_g, testing w_g=$\{1, 2\}$ to explore the effect of enforcing word boundaries in the grow-diag process.

5.1 English-Finnish results

We test the translation quality of both directions (*en-fi*) and (*fi-en*). We present results in Table 4 for 7 systems, including: our baselines, three CCM-based systems with word-boundary knowledge w_g=$\{0, 1, 2\}$ and two w_m-systems w_g=$\{1, 2\}$.

Results in Table 4 show that our CCM approach effectively improves the performance. Compared to the w_m-system, it chalks up a gain of 0.46 BLEU points for *en-fi*, and a larger improvement of 0.93 points for the easier, reverse direction.

Further using word boundary knowledge in our modified grow-diag process demonstrates that the additional flexibility consistently enhances BLEU for $w_g = 1, 2$. We achieve the best performance at $w_g = 2$ with improvements of 0.67 and 1.22 BLEU points for *en-fi* and *fi-en*, respectively.

	en-fi	fi-en
w-system	14.58	23.56
w_m-system	14.47	22.89
w_m-system + CCM	$14.93_{+0.46}$	$23.82_{+0.93}$
w_m-system + CCM + $w_g = 1$	15.01	23.95
w_m-system + CCM + $w_g = 2$	$\mathbf{15.14}_{+0.67}$	$\mathbf{24.11}_{+1.22}$
w_m-system + $w_g = 1$	14.44	22.92
w_m-system + $w_g = 2$	14.28	23.01
(Ganchev, 2008) - Base	14.72	22.78
(Ganchev, 2008) - Postcat	14.74	$23.43_{+0.65}$
(Yang, 2006) - Base	N/A	22.0
(Yang, 2006) - Backoff	N/A	$22.3_{+0.3}$

Table 4: **English/Finnish results.** Shown are BLEU scores (in %) with subscripts indicating absolute improvements with respect to the w_m-system baseline.

Interestingly, employing the word boundary heuristic alone in the original grow-diag does not yield any improvement for *en-fi*, and even worsens as w_g is enlarged (as seen in Rows 6–7). There are only slight improvements for *fi-en* with larger w_g.This attests to the importance of combining the CCM model and the modified grow-diag process.

Our best system outperforms the w-system baseline by 0.56 BLEU points for *en-fi*, and yields an improvement of 0.55 points for *fi-en*.

Compared to works experimenting *en/fi* translation, we note the two prominent ones by Yang and Kirchhoff (2006) and recently by Ganchev et al. (2008). The former uses a simple back-off method experimenting only *fi-en*, yielding an improvement of 0.3 BLEU points. Work in the op-

posite direction (*en-fi*) is rare, with the latter paper extending the EM algorithm using posterior constraints, but showing no improvement; for *fi-en*, they demonstrate a gain of 0.65 points. Our CCM method compares favorably against both approaches, which use the same datasets as ours.

5.2 English-Hungarian results

To validate our CCM method as language-independent, we also perform preliminary experiments on *en-hu*. Table 5 shows the results using the same CCM setting and experimental schemes as in *en/fi*. An improvement of 0.35 BLEU points is shown using the CCM model. We further improve by 0.44 points with word boundary w_g=1, but performance degrades for the larger window. Due to time constraints, we leave experiments for the reverse, easier direction as future work. Though modest, the best improvement for *en-hu* is statistical significant at $p<0.01$ according to Collins' sign test (Collins et al., 2005).

System	BLEU
w-system	9.63
w_m-system	9.47
w_m-system + CCM	9.82 $_{+0.35}$
w_m-system + CCM + $w_g = 1$	**9.91** $_{+0.44}$
w_m-system + CCM + $w_g = 2$	9.87

Table 5: **English/Hungarian results.** Subscripts indicate absolute improvements with respect to the w_m-system.

We note that MT experiments for *en/hu* [4] are very limited, especially for the *en* to *hu* direction. Novák (2009) obtained an improvement of 0.22 BLEU with no distortion penalty; whereas Koehn and Haddow (2009) enhanced by 0.5 points using monotone-at-punctuation reordering, minimum Bayes risk and larger beam size decoding.

While not directly comparable in the exact settings, these systems share the same data source and splits similar to ours. In view of these community results, we conclude that our CCM model does perform competitively in the *en-hu* task, and indeed seems to be language independent.

6 Detailed Analysis

The macroscopic evaluation validates our approach as improving BLEU over both baselines, but how do the various components contribute? We first analyze the effects of Step 4 in producing the CCM alignment, and then step backward to examine the contribution of the different feature classes in Step 3 towards the ME model.

6.1 Quality of CCM alignment

To evaluate the quality of the predicted CCM alignment, we address the following questions:

Q1: What is the portion of CCM pairs being misaligned in the IBM Model 4 alignment?

Q2: How does the CCM alignment differ from the IBM Model 4 alignment?

Q3: To what extent do the new links introduced by our CCM model address Q1?

Given that we do not have linguistic expertise in Finnish or Hungarian, it is not possible to exhaustively list all misaligned CCM pairs in the IBM Model 4 alignment. As such, we need to find other form of approximation in order to address Q1.

We observe that correct links that do not exist in the original alignment could be entirely missing, or mistakenly aligned to neighboring words. With morpheme input, we can also classify mistakes with respect to intra- or inter-word errors. Figure 2 characterizes errors T_1, T_2 and T_3, each being a more severe error class than the previous. Focusing on e_i in the figure, links connecting e_i to $f_{j'}$ or $f_{j''}$ are deemed T_1 errors (misalignments happen on one side). A T_2 error aligns f''_j within the same word, while a T_3 error aligns it outside the current word but still within its neighborhood. This characterization is automatic, cheap and has the advantage of being language-independent.

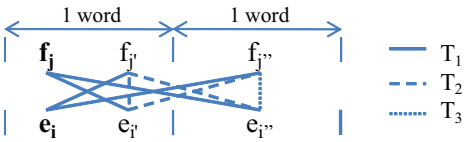

Figure 2: **Categorization of CCM missing links.** Given that a CCM pair link (f_j–e_i) is not present in the IBM Model 4, occurrences of any nearby link of the types $T_{[1-3]}$ can be construed as evidence of a potential misalignment.

Statistics in Table 6(ii) answers Q1, suggesting a fairly large number of missing CCM links: $3,418K$ for *en/fi* and $6,216K$ for *en/hu*, about 12.35% and 12.06% of the IBM Model 4 union alignment respectively. We see that T_1 errors con-

[4] Hungarian was used in the ACL shared task 2008, 2009.

stitute the majority, a reasonable reflection of the *garbage- collector*[5] effect discussed in Section 3.

	General (i)		Missing CCM links (ii)		
	en/fi	en/hu		en/fi	en/hu
Direct	17,632K	34,312K	T_1	2,215K	3,487K
Inverse	18,681K	34,676K	T_2	358K	690K
$D \cap I$	8,643K	17,441K	T_3	845K	2,039K
$D \cup I$	27,670K	51,547K	**Total**	**3,418K**	**6,216K**

Table 6: **IBM Model 4 alignment statistics.** (i) General statistics. (ii) Potentially missing CCM links.

Q2 is addressed by the last column in Table 7. Our CCM model produces about 11.98% (1,035K/8,643K) new CCM links as compared to the size of the IBM Model 4 intersection alignment for en/fi, and similarly, 9.52% for en/hu.

	Orig.	Resolved	I	U\I	New
en/fi	5,299K	3,433K	1065K	1,332K	**1,035K**
en/hu	9,425K	6,558K	2,752K	2,146K	**1,660K**

Table 7: **CCM vs IBM Model 4 alignments.** *Orig.* and *Resolved* give # CCM links predicted in Step 4 before and after resolving conflicts. Also shown are the number of resolved links present in the **I**ntersection, Union excluding I (**U\I**) of the IBM Model 4 alignment and **New** CCM links.

Lastly, figures in Table 8 answer Q3, revealing that for *en/fi*, 91.11% (943K/1,035K) of the new CCM links effectively cover the missing CCM alignments, recovering 27.59% (943K/3,418K) of all missing CCM links. Our modified *grow-diag* realizes a majority 76.56% (722K/943K) of these links in the final alignment.

We obtain similar results in the *en/hu* pair for link recovery, but a smaller percentage 22.59% (330K/1,461K) are realized through the modified symmetrization. This partially explains why improvements are modest for *en/hu*.

	New CCM Links (i)		Modified grow-diag (ii)	
	en/fi	en/hu	en/fi	en/hu
T_1	707K	1,002K	547K	228K
T_2	108K	146K	79K	22K
T_3	128K	313K	96K	80K
Total	**943K**	**1,461K**	**722K**	**330K**

Table 8: **Quality of the newly introduced CCM links**. Shown are # new CCM links addressing the three error types before (i) and after (ii) the modified grow-diag process.

6.2 Contributions of ME Feature Classes

We also evaluate the effectiveness the ME features individually through ablation tests. For brevity, we only examine the more difficult translation direction, *en* to *fi*. Results in Table 9 suggest that all our features are effective, and that removing any feature class degrades performance. Balancing specificity and generality, `bi1` is the most influential feature in the bilingual context group. For monolingual context, `inter`, which captures larger monolingual context, outperforms `intra`. The most important feature overall is `pmi`, which captures global alignment preferences. As feature groups, bilingual and monolingual context features are important sources of information, as removing them drastically decreases system performance by 0.23 and 0.16 BLEU, respectively.

System			BLEU
all (w_m-system+CCM)			14.93
−bi2	14.90	−intra	14.89
−bi1	$14.84^{*}_{-0.09}$	−pmi	$14.81^{*}_{-0.12}$
−bi0	14.89	−bi{2/1/0}	$14.70^{*}_{-0.23}$
−inter	14.85	−in{ter/tra}	$14.77^{*}_{-0.16}$

Table 9: **ME feature ablation tests for English-Finnish experiments.** $*$ mark results statistically significant at $p < 0.05$, differences are subscripted.

7 Conclusion and Future Work

In this work, we have proposed a language-independent model that addresses morpheme alignment problems involving highly inflected languages. Our method is unsupervised, requiring no language specific information or resources, yet its improvement on BLEU is comparable to much semantically richer, language-specific work. As our approach deals only with input word alignment, any downstream modifications of the translation model also benefit.

As alignment is a central focus in this work, we plan to extend our work over different and multiple input alignments. We also feel that better methods for the incorporation of CCM alignments is an area for improvement. In the en/hu pair, a large proportion of discovered CCM links are discarded, in favor of spurious links from the union alignment. Automatic estimation of the correctness of our CCM alignments may improve end translation quality over our heuristic method.

[5] E.g., e_i prefers f'_j or f''_j (garbage collectors) over f_j.

References

Berger, Adam L., Stephen D. Della Pietra, and Vincent J. D. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Bouma, Gerlof. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference*, Tübingen, Gunter Narr Verlag.

Cherry, Colin and Dekang Lin. 2007. Inversion transduction grammar for joint phrasal translation modeling. In *SSST*.

Collins, Michael, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *ACL*.

Creutz, Mathias and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.*, 4(1):3.

DeNero, John and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *ACL*.

Ganchev, Kuzman, João V. Graça, and Ben Taskar. 2008. Better alignments = better translations? In *ACL-HLT*.

Goldwater, Sharon and David McClosky. 2005. Improving statistical mt through morphological analysis. In *HLT*.

Haghighi, Aria, John Blitzer, John DeNero, and Dan Klein. 2009. Better word alignments with supervised itg models. In *ACL*.

Koehn, Philipp and Barry Haddow. 2009. Edinburgh's submission to all tracks of the WMT2009 shared task with reordering and speed improvements to Moses. In *EACL*.

Koehn, Philipp, Hieu Hoang, Alexandra Birch Mayne, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL, Demonstration Session*.

Lee, Young-Suk. 2004. Morphological analysis for statistical machine translation. In *HLT-NAACL*.

Liang, Percy, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *HLT-NAACL*.

Matusov, Evgeny, Richard Zens, and Hermann Ney. 2004. Symmetric word alignments for statistical machine translation. In *COLING*.

Moore, Robert C. 2004. Improving IBM word-alignment model 1. In *ACL*.

Nguyen, Thuy Linh and Stephan Vogel. 2008. Context-based Arabic morphological analysis for machine translation. In *CoNLL*.

Novák, Attila. 2009. MorphoLogic's submission for the WMT 2009 shared task. In *EACL*.

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

Och, Franz Josef and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Setiawan, Hendra, Min-Yen Kan, and Haizhou Li. 2007. Ordering phrases with function words. In *ACL*.

Virpioja, Sami, Jaakko J. Vyrynen, Mathias Creutz, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *MT Summit XI*.

Yang, Mei and Katrin Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *EACL*.

Zhang, Hao and Daniel Gildea. 2005. Stochastic lexicalized inversion transduction grammar for alignment. In *ACL*.

Zhang, Hao, Chris Quirk, Robert C. Moore, and Daniel Gildea. 2008. Bayesian learning of non-compositional phrases with synchronous parsing. In *ACL-HLT*.