Morphology-Aware Alignments for Translation to and from a Synthetic Language

Franck Burlot, François Yvon

LIMSI, CNRS, Université Paris Saclay, 91 403 Orsay, France

firstname.lastname@limsi.fr

Abstract

Most statistical translation models rely on the unsupervized computation of word-based alignments, which both serve to identify elementary translation units and to uncover hidden translation derivations. It is widely acknowledged that such alignments can only be reliably established for languages that share a sufficiently close notion of a word. When this is not the case, the usual method is to pre-process the data so as to balance the number of tokens on both sides of the corpus. In this paper, we propose a *factored alignment model* specifically designed to handle alignments involving a synthetic language (using the case of the Czech:English language pair). We show that this model can greatly reduce the number of non-aligned words on the English side, yielding more compact translation models, with little impact on the translation quality in our testing conditions.

1. Introduction

Most statistical translation models rely on the unsupervized computation of word-based alignments, which serve both to identify elementary translation units, as in phrase-based [1] and hierarchical [2] Machine Translation (MT) and to uncover hidden translation derivations, as in n-gram-based MT [3]. The *de-facto* standard for computing such alignments is to use the IBM models [4], as implemented in efficient software packages such as GIZA++ [5, 6] or `fast_align` [7].

It is however widely acknowledged that such alignments can only be reliably established for languages that share a sufficiently close notion of a word. When this is not the case, the usual method is to pre-process the data so as to balance the number of tokens on both sides of the corpus. Assuming translation into English from a morphologically rich language, this process will decompose complex source forms into shorter segments, and/or neutralize morphological variations that are not overly marked (and thus not necessary for the translation process) in the morphologically simpler one: forms that only differ in their case marking can, for instance, be collapsed into one non-marked version for the purpose of translating into English. This situation also occurs, though in a more extreme form, when translating from a language without explicit word separators such as Chinese [8, 9].

This strategy has been successfully applied to many language pairs in the context of MT applications: [10] is a first attempt to cluster morphological variants when translating from German into English; while [11] focuses on splitting German compounds. Similar techniques have been proposed for other language pairs such as Czech [12], Arabic [13, 14], Spanish [15], Finnish [16], Turkish [17] to name a few early studies. Note that the benefits (in terms of translation quality) of such pre-processing can be limited, except for the translation of out-of-vocabulary words.

In this paper, we focus on a slightly different problem, which arises when aligning English with a synthetic language. In this situation, many English words, notably function words such as determiners, pronouns and prepositions, may have no direct equivalent on the source side, in cases where for example their function is expressed morphologically by bound morphemes. Such problems, and their detrimental consequences for MT, are more thoroughly discussed in § 2 taking the Czech:English language pair as the main source of examples. To mitigate this undesirable situation, we propose a *factored alignment model* specifically designed to handle alignments involving a synthetic language, (see § 3, where we introduce these new variants of IBM Model 2). In our experiments with MT from and into English (§ 4), we show that this model can greatly reduce the number of non-aligned words on the English side, yielding more compact translation models, with little impact on the translation quality in our testing conditions. We finally discuss related work (§ 5) and conclude with further prospects.

2. Alignments with a Synthetic Language

Czech is a morphologically rich language with complex nominal, adjectival and verbal inflection systems. For instance, compared to the English adjective, which is invariable, its Czech counterpart has many different forms, varying in case (7), number (2) and gender (3). Therefore, Czech words contain more information than in English, which is typical of a synthetic language. On the other hand, the same kind of information may be encoded in a separate word in English, a language that has analytical tendencies. For instance, the Czech nominal genitive marker plays a similar role to the English preposition *of*, as in *the engine **of** the car* $\rightarrow$ *motor aut**a***.

Therefore, when aligning those two languages, linking a Czech noun (or verb, or adjective) solely to its English counterpart is quite unsatisfactory, since the information encoded in the Czech word ending is then lost (see Figure 1);

Table 1: Unaligned preposition causing a mistake (Czech-English).

source	Na seznamu jsou **v první řadě** plány na rozsáhlejší spolupráci v oblasti jaderné energetiky.
output	On the list are the **first in a series of** plans for greater cooperation in the field of nuclear energy.
ref.	High on the agenda are plans for greater nuclear co-operation.

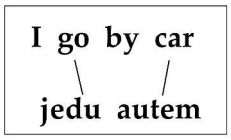


Figure 1: Lexical alignments missing the English pronoun and preposition that are encoded in the Czech endings.

and it might be desirable to also align neighboring function words on the English side. Missing these links indeed leads to mistakes in the output. In the Moses [18] baseline for Czech to English described in § 4, we often observed that an unaligned English preposition is associated to the wrong phrase, leading to a translation error, as illustrated in Table 1. In this example, the Czech *v první řadě* means literally *in first-Locative rank-Locative* and the phrases that were selected incorrectly include prepositions that were not aligned:

- **v první - first in**: this phrase pair leaves out the translation of the Czech preposition *v* and includes an English preposition that has no equivalent in the source, and might be erronously aligned to *v*.
- **řadě - a series of**: the Czech locative case is not translated and the English preposition *of* is not present on the Czech side.

We observe that standard alignment toolkits tend to miss such links. Table 2 reports the ratio of English words that remained unaligned after we trained alignments in both directions with symmetrization, using `fast_align`. Among the 7% unaligned words, almost 50% are determiners, which was predictable, since Czech does not have articles. Prepositions account for 33.2% of the unaligned words, over 10 points more than what we observe when aligning French and English. A similar situation happens with Russian, where more than 20% of English prepositions have no alignment. This suggests a difference between languages with synthetic tendencies such as Czech or Russian and more analytical ones such as English in the way they encode grammatical features such as case. When running asymmetric alignments from Czech to English, numbers are even worse, with 52.9% of the English prepositions remaining unaligned. We conclude that there is often no preposition on the Czech side to be linked to an English one. On the contrary, aligning French or Spanish to English means fewer unlinked prepositions and a higher rate of unaligned nouns. Hence, the problem of function word alignments is less obvious and the information we lose the most is lexical, rather than grammatical.

We argue that a more suitable alignment should extract phrases in which the English preposition is more systematically co-aligned with its head noun. This would make the extraction of phrases with a dangling, unaligned *of* less likely, and contribute to fixing certain case agreement errors.

Unaligned words are not only a problem in terms of the translation of prepositions. Since Czech is a pro-drop language, many English subject personal pronouns have no source to align to, leading to their omissions in many hypothesis translations when translating into English, such as in the clause with no subject found in one of the outputs of our baseline systems *and will go into it*. Aligning more English pronouns to Czech verbs should help to capture the necessity of jointly translating a verb into a pronoun and a verb in the target. In our English-to-Czech baseline (§ 4), we also often encounter situations where a negative Czech verb is translated into an affirmative form in English. Since Czech negation is encoded as a prefix (*ne-*, see Table 3), it is difficult to align it to English words such as *not*.[1]

Note that the units we need to find alignments for on the Czech side always encode grammatical information: person, negation and case, which should align to English function words. This is the main motivation for our proposal to add morphological alignments on top of lexical ones.

3. Morphological Alignment Model

3.1. Aligning words with feature vectors

Our model aims to make word-to-word alignments more dense by linking morphological tags on the Czech side to English function words. We first perform a morphological analysis of Czech and obtain a vector-based representation for each token, containing the lemma and various morphological labels (see § 2). Our model thus assumes sentences taking the form of a vector $\mathbf{e}$ of I word forms on the English side and of a $K \times J$ matrix $\mathbf{f}$ on the Czech side, where each row corresponds to various features of the word (such as lemma, person and case, as shown in Figure 2.a). By convention, we assume that the lemma is at index 1 in vector f_j.

Using these notations, our alignment model is a simple variant of IBM model 2 where (a) lemmas are aligned independently from one another, and (b) tag alignments are inde-

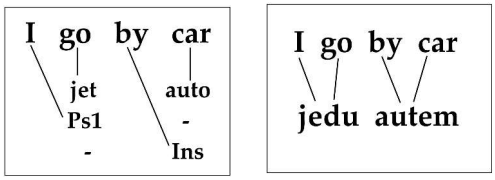
[1] The adverb *not* makes up the majority of unaligned adverbs in Table 2.

Table 2: Unaligned English words with symmetrized alignments across four language pairs using `fast_align`. $\frac{POS}{unali.}$: rate of unaligned occurrences of the POS over all unaligned words ; $\frac{unali.}{POS}$: rate of unaligned words over all occurrences of the POS.

POS	Cs-En (asym)		Cs-En (sym)		Ru-En		Fr-En		Es-En	
	$\frac{POS}{unali.}$	$\frac{unali.}{POS}$	$\frac{POS}{unali.}$	$\frac{unali.}{POS}$	$\frac{POS}{unali.}$	$\frac{unali.}{POS}$	$\frac{POS}{unali.}$	$\frac{unali.}{POS}$	$\frac{POS}{unali.}$	$\frac{unali.}{POS}$
Determiners	26.2%	65.2%	48.7%	30.1%	16.2%	31.0%	13.0%	11.6%	15.1%	4.4%
Prepositions	28.6%	52.9%	33.2%	15.3%	19.1%	23.3%	20.1%	12.4%	32.4%	7.2%
Auxiliaries	9.7%	37.6%	4.3%	4.4%	5.4%	19.5%	6.4%	11.8%	11.9%	5.6%
Nouns	8.7%	8.8%	3.4%	0.9%	26.7%	14.8%	28.6%	7.6%	8.1%	1.1%
Adverbs	4.9%	26.8%	1.9%	2.5%	3.6%	17.8%	3.2%	9.6%	6.3%	4.1%
Pers. Pronouns	7.3%	65.5%	0.6%	1.2%	2.5%	15.8%	1.6%	9.9%	3.0%	2.5%
Aligned words	72.0%		93.0%		81.6%		90.3%		96.3%	

Table 3: Unaligned negation adverb causing a mistake (English-Czech).

source	he is **not** at all aggressive
output	je vůbec agresivní
	he is at all aggressive
ref.	**ne**ní vůbec agresivní
	he is not at all aggressive



(a) (b)

Figure 2: Morphological alignments. (a) The source 1st person tag is aligned to the target pronoun *I* and the instrumental case tag to the preposition *by*. (b) Lemma and tag alignments are merged to provide links between word forms.

pendent given the alignment of their lemma, yielding:

$$p(\boldsymbol{f}|\boldsymbol{e}) = \sum_{\boldsymbol{a}} \prod_{j=1}^{J} \Big[p(a_{j1}|\boldsymbol{e}) p(f_{j1}|e_{a_{j1}}) \times \prod_{k=2}^{K} p(a_{jk}|a_{j1}) p(f_{jk}|e_{a_{jk}}) \Big] \quad (1)$$

This model thus allows us to integrate into the alignment probability the morphological properties of a lemma, which should for instance reinforce the alignment of a Czech noun with an English noun when the former is marked with a case that often matches a nearby preposition of the latter. Note that using the IBM model 2 is somewhat oversimplistic, as it assumes for instance that morphological markers of close words are unrelated, even though agreement rules enforce similar cases for words within the same noun phrase. A more realistic version, in which such dependencies would be modeled at least indirectly, would be to use a better distortion model to constrain the alignment of neighboring lemmas. Given the implementation choices described above, it was not necessary to develop this idea any further.

To complete the description, note that we assume that the alignment of the lemma (a_{j1}) only depends on j, I and J; and that the alignments of the morphological tags (a_{jk}) only depend on the difference ($a_{jk} - a_{j1}$). We further enforce $p(a_{jk}|a_{j1}) = 0$ outside of a fixed-size window centered on a_{j1} (3 words to the left side, one word to the right side).[2] The model defined in Equation (1) lends itself well to estimation via EM. We however also performed experiments with more constrained implementations, as described below.

3.2. Implementation variants

In the experiments reported below, we constrast various implementations of this alignment model in the computation of the Czech-to-English alignments; note that we use a standard word-based IBM model for the other direction. A first condition (joint//ibm in Table 10) uses a faithful implementation of EM for the model of Equation (1), in which we initialize uniformly the translation and the distortion parameters.

A second condition uses the output of a first pass alignment to better constraint the alignments of lemmas. The first stage computes alignments between Czech lemmas and English words using standard word alignment pipelines: in our experiments, we used both asymmetric alignments computed with IBM model 2 and IBM model 4, or symmetrized alignments obtained by running these models in both directions. In any case, we keep these alignment links fixed during the second stage, in which we estimate the morphological alignment model and compute alignments links for tags.

A softer version of the second condition is to use the first pass alignments to initialize the translation model, which are then free to change in the course of the EM procedure.

Finally note that we also enforce a void alignment for "null" morphological tags (eg. the case marking for verbs, or the tense of nouns, see Figure 2.b).

For all conditions, training involves multiples iterations

[2] As for the right side, we consider only one position to target words like *not* and *'s*, as in *can **not**, Hana **'s** hand*.

of EM with models of increasing complexity for a fixed number of iterations. We first train the lemma-to-word alignments, before also considering the tags-to-word parameters. A final decoding computes the optimal alignment for morphological tags; at this stage, we only keep alignment links that match a non-aligned word on the English side, and use these to complete the baseline alignment, as shown in Figure 2.b. The rest of the training of the translation model (phrase extraction, etc.) remains unchanged.

4. Experimental Results

4.1. Data and Experimental Setup

We used two datasets to train our SMT systems:

- **A small dataset** consisting of about 790k parallel sentences taken from the Europarl [19] and News Commentary corpora distributed for the shared translation task of WMT 2015.[3] The monolingual data is made up of one side of the parallel corpora and the News Crawl corpora (2014) and adds up to 29M sentences for English and 37M for Czech.

- **A bigger dataset** of about 15M parallel sentences, composed of the previous set and the Czeng 1.0 corpus [20]. We added to the monolingual data one side of the Czeng 1.0 corpus and the previous versions of the News Crawl corpora (2007-2013). and obtained a total of 52M Czech and 43M English sentences.

This data is tokenized and true-cased before starting the alignment. The morphological analysis on the Czech side is performed using MorphoDiTa [21]. After word alignment, all downstream training steps are carried out using the Moses toolkit [18]: this includes phrase extraction and scoring, lexical weighting and learning the lexicalized reordering models. 4-gram language models are trained with KenLM [22] for both languages. Tuning is performed using MERT [23] on the test set of the WMT 2014 translation task. For the sake of comparison, we also report results obtained with n-gram-based systems trained with Ncode [3, 24].

4.2. Alignment Setup

We used M morphological features to fill the Czech word vectors $\mathbf{f}$ in our experiments: case, person, time/mode, and whether a verb has a negative form - Czech representations have therefore $M = 5$ dimensions.

Regarding condition 1, where lexical alignments are learnt jointly with morphological links (for Czech-to-English), 4 strategies were tested:

- **ibm//none**: only forward (cs-en) alignments;

- **joint//none**: only forward (cs-en) alignments trained according to our model;

- **ibm//ibm**: forward and backward alignments symmetrized with the grow-diag-final-and heuristics;

- **joint//ibm**: symmetrization is performed with joint-none and the backward (en-cs) alignments;

Regarding the training condition 2, we used `fast_align` (resp. `Mgiza`) to get initial IBM2 (resp. IBM4) alignments between Czech lemmas and English words. We added to the former 3 strategies to obtain different alignment types:

- **ibm+morph//none**: forward and morphological alignments;

- **ibm+morph//ibm**: a symmetrized version also involving backward en:cz alignments;

- **[ibm//ibm]+morph**: morphological alignment is performed after symmetrization.

During decoding, the most likely morphological alignments are subject to three constraints in order to be accepted:

- The candidate English lemma should not be aligned;

- The morphological alignment probability should be higher than a threshold (0.05 in our experiments);

- The candidate English lemma should have a frequency higher than 1,000 occurrences (15,000 for the bigger data set) in the English part of the parallel corpus.

These heuristics help to improve the quality of alignment by reducing links with rare words that may have a high probability, given a specific tag. Since the words we target are mainly English function words (pronouns, prepositions, etc.), it seems reasonable to focus on a small set of high frequency tokens. Note finally that the same word alignments were used both to train the en-cs and the cs-en systems.

4.3. Results

Morphological alignments effectively address the problem of previously unaligned words by linking function words, as reflected in Table 4, even though ibm+morph//none also returns a few more alignments for nouns. This shows that some lexical alignments had also been wrongly performed, most of which are corrected by symmetrization in the ibm+morph//ibm variant. The first impact of morphological alignments is a reduction of the phrase table size: using `fast_align`, we lost almost 1.5M phrases when adding morphological alignments to the symmetrized baseline, meaning that over 6% of initial phrases have been discarded (see Table 5).[4] `Mgiza` alignements show the clearest contrast, since the number of phrase pairs for ibm//ibm (44M) is reduced to less than 28M in ibm+morph//ibm.

[3] `http://statmt.org/wmt15/`

[4] Note that if the number of phrase pairs is lower, the average length of phrases stay the same in every system. For instance, ibm//ibm has 3.77 tokens per Czech phrase and 4.26 per English one, which is very similar to [ibm//ibm]+morph with respectively 3.79 and 4.25 tokens per phrase.

Table 4: Links added by morphological alignments (Czech-English) using `fast_align`. $\frac{POS}{unali.}$: rate of unaligned occurrences of the POS over all unaligned words ; $\frac{unali.}{POS}$: rate of unaligned words over all occurrences of the POS.

POS	ibm//none		ibm+morph//none		ibm//ibm		[ibm//ibm]+morph		joint//ibm	
	$\frac{POS}{unali.}$	$\frac{unali.}{POS}$	$\frac{POS}{unali.}$	$\frac{unali.}{POS}$	$\frac{POS}{unali.}$	$\frac{unali.}{POS}$	$\frac{POS}{unali.}$	$\frac{unali.}{POS}$	$\frac{POS}{unali.}$	$\frac{unali.}{POS}$
Determiners	26.2%	65.2%	32.6%	58.2%	48.7%	30.1%	58.7%	28.5%	51.6%	24.3%
Prepositions	28.6%	52.9%	25.6%	34.0%	33.2%	15.3%	24.4%	8.8%	31.3%	11.0%
Auxiliaries	9.7%	37.6%	7.0%	20.6%	4.3%	4.4%	3.3%	2.7%	4.5%	3.5%
Nouns	8.7%	8.8%	9.4%	6.9%	3.4%	0.9%	3.4%	0.7%	3.0%	0.6%
Adverbs	4.9%	26.8%	5.0%	19.8%	1.9%	2.5%	2.0%	2.2%	1.9%	2.0%
Pers. Pronouns	7.3%	65.5%	4.7%	25.7%	0.6%	1.2%	0.7%	1.0%	0.7%	1.0%
Aligned words	72.0%		79.3%		93.0%		94.4%		94.6%	

Table 5: Results in BLEU for Czech-English (smaller data condition).

	`fast_align` (IBM2)			`Mgiza` (IBM4)	
Alignment Setup	Ncode	Moses	Phrase Table Size	Moses	Phrase Table Size
ibm//none	-	20.34	50,462,274	20.31	56,967,921
ibm+morph//none	-	19.98	35,364,892	20.26	45,549,682
ibm+morph//ibm	-	20.08	20,286,841	20.14	27,820,416
ibm//ibm	19.72	20.34	22,799,794	20.35	44,410,638
[ibm//ibm]+morph	19.68	20.26	21,247,701	20.33	40,805,062

Table 6: Results in BLEU for English-Czech (for the small data condition). The size of the phrase tables is the same as in Table 5.

	`fast_align`		`Mgiza`
Alignment Setup	Ncode	Moses	Moses
ibm//none	-	13.94	14.24
ibm+morph//none	-	13.90	14.03
ibm+morph//ibm	-	14.02	13.91
ibm//ibm	14.02	14.09	14.45
[ibm//ibm]+morph	14.03	14.21	14.20

Table 7: Better case prediction (English-Czech).

source	who are captured by Ukrainian soldiers
ibm//ibm	kteří zadržený ukrajinští vojáci
	who-Plur captured-Passive-Sing Ukrainian-Nom soldiers-Nom
[ibm//ibm]+morph	kteří jsou zajati ukrajinskými vojáky
	who-Plur are captured-Passive-Plur Ukrainian-Ins soldiers-Ins

We evaluated our systems using the test set of the WMT 2015 translation shared task. Even though the effect on the BLEU score is minor, we observe a slight improvement when translating into Czech with `fast_align`[5] (see Table 6), which is understandable, since case is the major morphological category ignored by baseline alignments. Thus the new phrase table helps to better predict case inflection, mainly according to the preposition in the source sentence. Indeed, Table 7 shows the wrong translation of the English preposition *by* in the ibm//ibm system where the noun phrase is in nominative case. Our [ibm//ibm]+morph system successfully translates the preposition by the instrumental case needed for such passive constructions. Moreover, in the same direction, handling negation also helped to fix some baseline system errors, as for the example in Table 3 (our system actually outputs the reference sentence).

Note that a better management of case is also beneficial in the inverse direction (Czech-English), as shown in Table 8, where the erroneous phrase pairs described in § 2 (*v první - first in*;) *řadě - a series of* get a lower probability, allowing the correct translation to be selected during decoding. As a result, we observe that the most frequent prepositions (*of, to, in, for*) are generated less often in [ibm//ibm]+morph (4,070) than in the ibm//ibm (4,190), which we interpret as a sign of more relevant use of English prepositions in a morphology-aware system.

For the same translation direction, the number of subject personal pronouns is higher in [ibm//ibm]+morph (1,629) than in ibm//ibm (1,561), which suggests better constructions in the English output, such as in Table 9, where the Czech verb with no subject expressed is translated by a verb with its subject pronoun corresponding to the source word ending.

Furthermore, handling negation during the alignment step also yields improvement when translating into English. Indeed, the word *not* has 206 occurrences in ibm//ibm and 234 in [ibm//ibm]+morph, suggesting that the latter system

[5]The descriptions of our outputs relate to the alignments performed using `fast_align`.

Table 8: Better preposition extraction for relevant phrases (Czech-English).

source	Na seznamu jsou **v první řadě** plány na rozsáhlejší spolupráci v oblasti jaderné energetiky.
ibm//ibm	On the list are the **first in a series of** plans for greater cooperation in the field of nuclear energy.
[ibm//ibm]+morph	On the list are **primarily** plans for greater cooperation in the field of nuclear energy .

Table 9: Subject personal pronoun generation (Czech-English).

source	a bude**me** si ho rozebírat and will-Ps1-Plur it analyse
ibm//ibm	and will go into it
[ibm//ibm]+morph	and **we** will discuss it

conveys negation more.

Alignments with the time and mode tags for verbs helped to generate more correct English analytical constructions: while ibm//ibm omits the auxiliary in the translation of a Czech present verb into a passive form (*who usually based*), [ibm//ibm]+morph generates the right construction, despite the insertion of an adverb between both verbs: *who **are** usually based*. Nevertheless, for 2,639 auxiliaries in the former, the latter contains 2,716 of them, bringing almost insignificant changes.

We notice slightly worse results with the condition 1, where joint//ibm is 1 BLEU point below ibm//ibm for Czech-English, and 0.6 for English-Czech (see Table 10). The number of phrase pairs is a lot lower here than with condition 2, since more alignments are generated, as is shown in Table 4. Nevertheless, the score of the joint//none systems in both directions show that these alignments are very noisy, since they greatly underperform the ibm//none system.

Finally, Table 11 suggests that no impact on the BLEU score compared to the baseline is to be expected using more data, while the total ratio of aligned words went from 91.7% to 93.6% and 7% of initial phrases were discarded from the table in [ibm//ibm]+morph.

Table 10: Results in BLEU with joint learning of morphological and lexical alignments using Moses for the small data condition (+`fast_align` init: parameter initialization with `fast_align` output)

Alignment Setup	cs-en	en-cs	Phrase Table Size
ibm//none	20.34	13.94	50,462,274
joint//none	18.69	13.05	31,482,262
ibm//ibm	20.34	14.09	22,799,794
joint//ibm	19.33	13.47	15,179,849
+ `fast_align` init	19.41	13.40	15,210,792

Table 11: Results in BLEU for the large data condition (`Mgiza` with Moses)

Alignment Setup	cs-en	en-cs	Phrase Table Size
ibm//ibm	24.04	16.48	324,969,903
[ibm//ibm]+morph	24.07	16.38	301,714,878

5. Related Work

Aligning English with "morphologically-complex" languages poses several challenges, depending on the exact differences between the source and target – it has, over the years, attracted a considerable amount of effort, which has only been briefly reviewed here. In fact, morphological complexity can have multiple consequences for alignment.

First, it is often assumed that the morphologically complex language has more word types, due for instance to a richer inflectional system: this is the case for French or Spanish, which have a much richer conjugation than English. This, in turn, yields sparser counts, and less reliable probability estimates for the alignment models (notwithstanding a high Out-of-Vocabulary (OOV) ratio at testing time). The simplest remedy is to normalize the target side, using lemmas or other kinds of abstraction instead of words for the purpose of the alignment [25, 26, 27]. Note that defining the optimal level of abstraction is not obvious and often requires a significant tuning effort. Going one step further, it may also be interesting to keep these abstract representations for translation, but this requires a non-trivial post-processing step to restore the correct inflection when translating *into* the morphologically rich language [28]. The alternative strategy, which translates word forms, is plagued with OOV issues and requires specific strategies to properly handle unknown forms - as in the factored-models approach of [29, 30]. In our own alignment model, we borrow the idea to compute a first-pass alignment based primarily on lemmas, which seems to be more effective than using full forms. However, in our case, morphological information is not used to smooth alignment counts, but rather to take account of the function words in the English side.

The other well documented issue with morphologically rich languages is that word forms are more complex, meaning that they are made of several parts (morphemes for basic lexical units, lexemes for compounds). Depending on the language under consideration, identifying the orthographical and/or phonological counterparts of this elementary units can be fairly easy (in the case of purely agglutinative languages) or near impossible (in the case of fusional languages), with a large number of in-between situations. Many rule-based attempts at performing such decompositions as a pre-processing of the source side text have nonetheless been entertained: see [12], Arabic [13, 14], Spanish [15], Finnish [16], Turkish [17] to cite a few. Note that the opposite approach, consisting of "splicing" English words into artificially complex forms has also been considered (eg. in [31]).

As noted by several authors, decomposing word forms into morphemes goes against the main intuition of phrase-based SMT, which favors the translation of large units, and it also reduces the effectiveness of language models, as it decreases the size of the context. To mitigate these potentially negative effects, it is possible to simultaneously consider multiple decomposition schemes, which are then recombined using system combination techniques [32, 33, 34]. This however requires mechanisms to generate multiple morphological decompositions of the same text, using for instance the unsupervized segmentation models of [35, 36, 37]. As pointed out in [38], performing morphological segmentation of the source independently of the target is vastly sub-optimal, and joint models for alignment and segmentations are probably more appropriate in a MT context eg. [38, 39]. Our main focus being a fusional language, we have not made any attempt to segment the source words into smaller morphemes, and have instead used a feature-based representation associating a lemma and morphological properties.

6. Conclusions

This paper has described a factored alignment model specifically designed to handle alignments involving a language with synthetic tendencies, such as Czech. We have shown that this model can greatly reduce the number of non-aligned words on the English side, yielding more compact translation models that contain more relevant phrases. Case is the morphological feature that produces most alignments, which turned out to give some improvement when translating into Czech. On the other hand, using time and mode did not bring the expected gain, although it did help to better translate verb inflection in Czech and constructions in English.

The reported improvement over the baseline systems is not confirmed by a straight BLEU improvement. However we showed that one-to-many alignments from Czech to English help to better take into account the specificities of each language. While the English output has more words than in the baseline system, such as negative adverbs, auxiliaries, pronouns (disregarding the fact that it has fewer prepositions), the Czech output is more concise, showing eg. fewer incorrect verbal constructions and more reliance on inflection, which leads to better agreement.

In future work, we intend to confirm these tendencies by (a) using an improved model of morphological alignments, with an improved modeling of the dependency between tags and lemmas, and (b) testing our model with other translation tasks involving a synthetic target language.

7. Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments and suggestions. This work has been partly funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 645452 (QT21).

8. References

[1] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proc. HLT–NAACL*, 2003, pp. 127–133.

[2] D. Chiang, "A hierarchical phrase-based model for statistical machine translation," in *Proc. ACL*, Ann Arbor, MI, 2005, pp. 263–270.

[3] J. B. Mariño, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. Fonollosa, and M. R. Costa-Jussà, "N-gram-based machine translation," *Computational Linguistics*, vol. 32, no. 4, pp. 527–549, 2006.

[4] P. F. Brown, J. Cocke, S. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," vol. 16, no. 2, pp. 79–85, 1990.

[5] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[6] Q. Gao and S. Vogel, "Parallel implementations of word alignment tool," in *Proc. SETQA–NLP*, 2008, pp. 49–57.

[7] C. Dyer, V. Chahuneau, and N. A. Smith, "A Simple, Fast, and Effective Reparameterization of IBM Model 2," in *Proc. NAACL*, Atlanta, Georgia, 2013, pp. 644–648.

[8] P.-C. Chang, M. Galley, and C. D. Manning, "Optimizing Chinese word segmentation for machine translation performance," in *Proc. WMT*, Columbus, Ohio, 2008, pp. 224–232.

[9] T. Chung and D. Gildea, "Unsupervised tokenization for machine translation," in *Proc. EMNLP*, Singapore, 2009, pp. 718–726.

[10] S. Nießen and H. Ney, "Toward hierarchical models for statistical machine translation of inflected languages," in *Proc. of the ACL 2001 Workshop on Data-Driven Methods in MT*, Toulouse, France, 2001, pp. 47–51.

[11] P. Koehn and K. Knight, "Empirical methods for compound splitting," in *Proc. EACL*, Budapest, Hungary, 2003, pp. 187–193.

[12] S. Goldwater and D. McClosky, "Improving statistical MT through morphological analysis," in *Proc. HLT–EMNLP*, Vancouver, Canada, 2005, pp. 676–683.

[13] Y.-S. Lee, "Morphological analysis for statistical machine translation," in *Proc. HLT-NAACL 2004: Short Papers*, 2004, pp. 57–60.

[14] F. Sadat and N. Habash, "Combination of arabic preprocessing schemes for statistical machine translation," in *Proc. COLING/ACL*, 2006, pp. 1–8.

[15] A. de Gispert and J. B. Mariño, "On the impact of morphology in English to Spanish statistical MT," *Speech Communication*, vol. 50, no. 11-12, pp. 1034–1046, 2008.

[16] S. Virpioja, J. J. Väyrynen, M. Creutz, and M. Sadeniemi, "Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner," in *Proc. MT Summit XI*, Copenhagen, Denmark, 2007, pp. 491–498.

[17] K. Oflazer and I. D. El-Kahlout, "Exploring different representational units in English-to-Turkish statistical machine translation," in *Proc. WMT*, 2007, pp. 25–32.

[18] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proc. ACL:Systems Demos*, Prague, Czech Republic, 2007.

[19] P. Koehn, "A parallel corpus for statistical machine translation," in *Proc. MT-Summit*, Phuket, Thailand, 2005.

[20] O. Bojar, Z. Žabokrtský, O. Dušek, P. Galuščáková, M. Majliš, D. Mareček, J. Maršík, M. Novák, M. Popel, and A. Tamchyna, "The Joy of Parallelism with CzEng 1.0," in *Proc. LREC2012*, ELRA. Istanbul, Turkey: ELRA, 2012.

[21] J. Straková, M. Straka, and J. Hajič, "Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition," in *Proc. ACL: System Demos*, Baltimore, Maryland, 2014, pp. 13–18.

[22] K. Heafield, "KenLM: Faster and Smaller Language Model Queries," in *Proc. WMT*, Edinburgh, Scotland, 2011, pp. 187–197.

[23] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proc. ACL*, 2003, pp. 160–167.

[24] J. M. Crego, F. Yvon, and J. B. Mariño, "N-code: an open-source Bilingual N-gram SMT Toolkit," *Prague Bulletin of Mathematical Linguistics*, vol. 96, pp. 49–58, 2011.

[25] H. Ney and M. Popovic, "Improving word alignment quality using morpho-syntactic information," in *Proc. COLING*, Geneva, Switzerland, 2004, pp. 310–314.

[26] A. de Gispert, D. Gupta, M. Popović, P. Lambert, J. Mariño, M. Federico, H. Ney, and R. Banchs, "Improving statistical word alignments with morpho-syntactic transformations," in *Advances in Natural Language Processing*, T. Salakoski, F. Ginter, S. Pyysalo, and T. Pahikkala, Eds. Springer Berlin Heidelberg, 2006, vol. 4139, pp. 368–379.

[27] M. Carpuat, "Toward using morphology in French-English phrase-based SMT," in *Proc. WMT*, Athens, Greece, 2009, pp. 150–154.

[28] A. Fraser, M. Weller, A. Cahill, and F. Cap, "Modeling inflection and word-formation in SMT," in *Proc. EACL*, Avignon, France, 2012, pp. 664–674.

[29] P. Koehn and H. Hoang, "Factored translation models," in *Proc. EMNLP–CoNLL*, Prague, Czech Republic, 2007, pp. 868–876.

[30] O. Bojar, "English-to-Czech factored machine translation," in *Proc. of the 2nd WMT*, Prague, Czech Republic, 2007, pp. 232–239.

[31] N. Ueffing and H. Ney, "Using POS information for statistical machine translation into morphologically rich languages," in *Proc. EACL*, Budapest, Hungary, 2003, pp. 347–354.

[32] C. J. Dyer, "The "noisier channel": Translation from morphologically complex languages," in *Proc. WMT*, Prague, Czech Republic, 2007, pp. 207–211.

[33] A. de Gispert, S. Virpioja, M. Kurimo, and W. Byrne, "Minimum Bayes Risk Combination of Translation Hypotheses from Alternative Morphological Decompositions," in *Proc. NAACL–HLT*, Boulder, Colorado, 2009, pp. 73–76.

[34] S. Virpioja, J. Väyrynen, A. Mansikkaniemi, and M. Kurimo, "Applying morphological decompositions to statistical machine translation," in *Proc. WMT and MetricsMATR*, Uppsala, Sweden, 2010, pp. 195–200.

[35] M. Creutz and K. Lagus, "Unsupervised models for morpheme segmentation and morphology learning," *ACM Trans. Speech Lang. Process.*, vol. 4, no. 1, pp. 3:1–3:34, Feb. 2007.

[36] S. Goldwater, T. L. Griffiths, and M. Johnson, "A Bayesian framework for word segmentation: Exploring the effects of context," *Cognition*, vol. 112, no. 1, pp. 21–54, 2009.

[37] D. Mochihashi, T. Yamada, and N. Ueda, "Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling," in *Proc. ACL/IJCNLP*, 2009, pp. 100–108.

[38] T. Nguyen, S. Vogel, and N. A. Smith, "Nonparametric word segmentation for machine translation," in *Proc. COLING*, Beijing, China, 2010, pp. 815–823.

[39] J. Naradowsky and K. Toutanova, "Unsupervised Bilingual Morpheme Segmentation and Alignment with Context-rich Hidden Semi-Markov Models," in *Proc. ACL*, Portland, OR, 2011, pp. 895–904.