Chapter 1

Related Work

1.1 Statistical machine translation

In this section, we review different machine translation *SMT* models as classified in Figure 1.1.

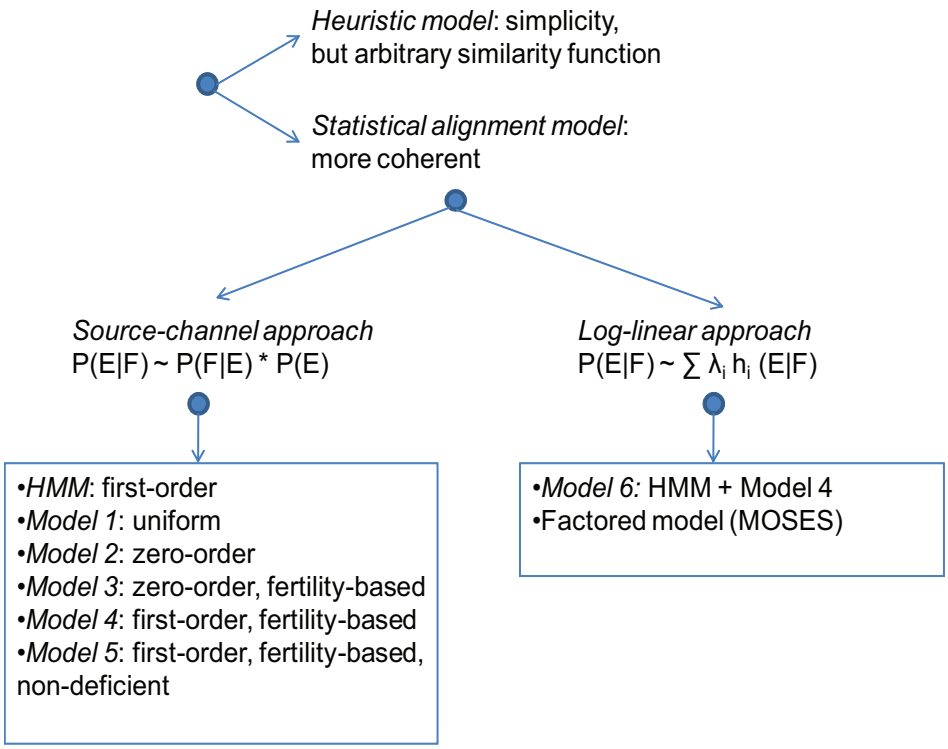

Figure 1.1: SMT model classification

Machine translation first started with heuristic models with the advantage of simplicity, in which similarity function is mainly used by capturing the co-occurrence of words. From the

association score matrix, suitable heuristics could be applied to derive word alignment, such as the following method in which Dice coefficient is used:

$$a_j = \arg\max_i dice(i, j)$$

According to (Och & Ney, 2003), one major problem of the heuristic models is the use of similarity functions which seems to be completely arbitrary. In view of that, more coherent class of models which is based on statistical alignment approach, is considered to be more appropriate. The statistical approach was originated from (Brown, Pietra, Pietra, & Mercer, 1993)with their influential paper proposing 5 IBM models. The statistical approach obtains association score by using statistical estimation theory, and model parameters are determined to maximize the model likelihood on the training corpus. Those 5 IBM models are designed under the unified view of *source-channel approach* in which a target language (E) goes through a noisy channel, and becomes the source langue (F), and the issue is to recover E from its distorted version, which is F. Follow this approach, HMM model and its extensions are suggested by (Vogel, Ney, & Tillmann, 1996) and (Och & Ney, 2003) to capture the locality property of word alignments. Recently, there emerges a more general promising approach, called *log-linear models*, which encapsulates the source-channel approach while allowing other features to be added in order to improve system performance. The log-linear approach was suggested by (Och & Ney, 2002) which contains the widely used source-channel approach as a special case. In (Och & Ney, 2003), a model called Model 6 is proposed which combines HMM model and IBM Model 4 in a log-linear way which yields significantly better results than simple heuristic models. Factored translation model (Koehn & Hoang, 2007) is recently proposed in which each word and its addition annotations (such as lemma, POS, morphology, or word class) are considered as a whole instead of the word alone. Together with factored translation model, a publicly-available SMT toolkit, MOSES (Koehn, Hoang, Mayne, Callison-Burch, Federico, Bertoldi, Cowan, Shen, Moran, Zens, Dyer, Bojar, Constantin, & Herbst, 2007b), is provided, and considered to be the state-of-the-art SMT system currently.

1.1.1 Source-channel approach

Source-channel approach in SMT is adopted from the analogy of Bayesian noisy channel models used in different applications such as speech or spelling. As illustrated in figure **bug figure**, when translating from a source sentence F to a target sentence E, we imagine that E has been distorted by the noisy channel to become F, and the translation process is to decode from F the most probable $\widehat{E}$. Under this approach, as translation task is described through the formula

$$\widehat{E} = arg\max_{E} Pr(F|E)\ .\ Pr(E) \tag{1.1}$$

in which Pr(F | E) represents translation model and Pr(E) means language model. Pr(F | E) tells how well a set of words in English (*target*) could be a translation of the French sentence (*source*). Pr(E), on the other hand, constrains on how well a sequence of word could be a good sentence in English. As such, Pr(E) helps the system on the linguistic aspects of the target language, e.g. imposing word ordering, deciding better word choice in translation, etc., which, in turn, relieves Pr(F | E) from many language-dependent issues to focus on finding goods set of words as candidate translations. For each pair of sentence (f_1^J, e_1^I), the translation probability could be expressed as the accumulated probability through different alignment possibilities a_1^J

$$Pr(f_1^J|e_1^I) = \sum_{a} Pr(f_1^J, a_1^J|e_1^I)$$

SMT models are generally different in representing $Pr(f_1^J, a_1^J|e_1^I)$, which in this section, we will highlight key differences among them as illustrated in Figure 1.1. We refer interested reader to other comprehensive surveys on SMT systems available in the literature such as (Och & Ney, 2003), or (Jurafsky & Martin, 2007). With respect to the aforementioned formula, the translation probabilities in HMM model, IBM Model 1 and Model 2 are abstractly defined as

$$Pr(f_1^J, a_1^J|e_1^I) = \text{length\_prob} * \text{alignment\_prob} * \text{lexicon\_prob}$$

where *length_prob* tells how likely the translated sentence will have a particular length, *alignment_prob* worries about the occurrence position of each translated word, and finally, *lexicon_prob* deals with the actual content of a translation, i.e. what should be translated from a

source word [1]. IBM Model 1 and Model 2 are considered zero-order dependencies (Och & Ney, 2003) where the alignment_prob of the target word j^{th} $p(a_j|j, I, J)$ does not depend on that of any preceding word. More specifically, IBM Model 1 uses an uniform distribution to assign $p(a_j|j, I, J) = 1/(I+1)$. HMM model, on the other hand, is first-order dependency in which the alignment_prob is expressed as $p(a_j|a_{j-1}, I)$. HMM model was designed to capture the strong *localization effect* that words are not distributed arbitrary over sentence positions, but tend to form cluster (Vogel et al., 1996).

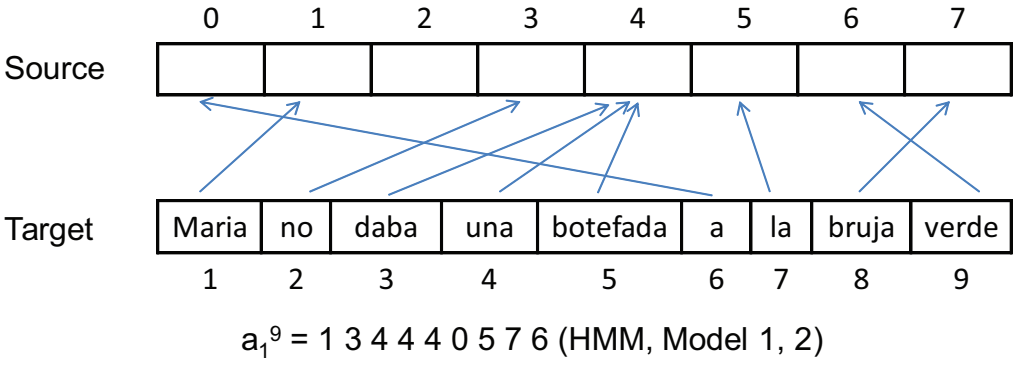

Figure 1.2: Alignment representation in HMM model, IBM Model 1, and 2

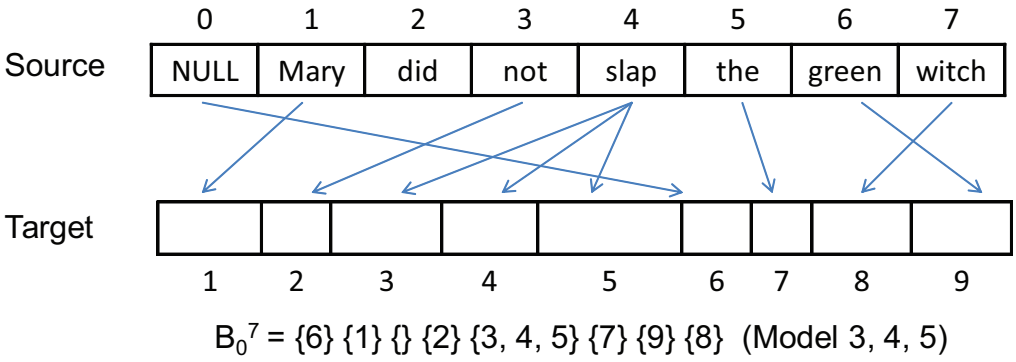

Figure 1.3: Alignment representation in IBM Model 3, 4, and 5

For IBM Model 3, Model 4, and Model 5, the representation is markedly different from the previous representation, and much more complicated due to the introduction of the *fertility* notions. Fertility of a source word is introduced to explicitly capture the number of target words that will be a translation for that source word, e.g. when slap is translated into "daba

[1] In the context of $Pr(f_1^J, a_1^J|e_1^I)$, we refer to e_1^I as source words, and f_1^J as target words

una botefada", we say that the fertility $\phi(slap) = 3$. With the notion of fertility, the perspective of an alignment is *reversed*. Instead of asking what position a_j in the source sentence corresponds to the target word f_j, we would like to know the set of target positions B_i that the source word e_i will translate to (see Figure 1.2 and 1.3). According to (Och & Ney, 2003), the translation probability for IBM Model 3, 4, and 5 $Pr(f_1^J, a_1^J|e_1^I)$, or equivalently $Pr(f_1^J, B_0^I|e_1^I)$, could be expressed as

$$Pr(f_1^J, a_1^J|e_1^I) = \text{spurious\_prob} * \text{alignment\_prob} * \text{lexicon\_prob}$$

in which alignment_prob denotes the probability of deciding B_i of each source word e_i, spurious_prob takes care of translating positions for the NULL source word B_0, and lexicon_prob concerns about translating to f_j given e_i. The 3 models differ at how they define alignment_prob. In IBM Model 3, the dependence of B_i on its predecessor B_{i-1} is ignored, i.e. zero-order dependency, while IBM Model 4 has first-order dependency. Both IBM Model 3 and 4 suffer from what is defined in (Brown et al., 1993) as *deficiency*, in which no constraint is imposed on alignment positions in B_i, for example, the positions may overlap. IBM Model 5 overcomes that weaknesses by only allow word e_i to choose vacant positions remained after words e_1^{i-1} have decided their translating positions. We end this section by providing Knight's nice illustration (Knight, 1999) on the generative process of IBM Model 3 in Figure 1.4. The process consists of five stages: input, choose fertilities, choose number of spurious words, choose translation, and choose target positions.

Mary did not slap the green witch (input)
Mary not slap slap slap the green witch (choose fertilities, e.g. $\phi(slap) = 3$)
Mary not slap slap slap NULL the green witch (choose number of spurious words)
Maria no daba una botefada a la verde bruja (choose translations)
Maria no daba una botefada a la bruja verde (choose target positions)

Figure 1.4: String-rewriting illustration for the generative process of IBM Model 3 (Knight, 1999)

1.1.2 Log-linear approach

In previous section, we have discussed about the source-channel approach, which indirectly solve the translation problem from *source to target* languages by considering the target language model as well as the translation model from *target to source* languages. An alternative to this approach is to model the posterior probability $Pr(E|F)$ directly (Och & Ney, 2002), and is referred as the *log-linear approach*. The linearity of the approach comes from the ability to combine M feature functions $h_m(E,F), m = 1 \ldots M$, contributing linearly to the posterior probability as

$$Pr(E|F) = \frac{exp \sum_{m=1}^{M} \lambda_m h_m(E,F)}{\text{normalizing\_factor}} \tag{1.2}$$

By taking the log on both side, our translation problem could be compactly represented as

$$\widehat{E} = arg \max_{E} \sum_{m=1}^{M} \lambda_m h_m(E,F)$$

The interesting characteristic of the log-linear approach is that it encapsulates our previous source-channel approach. Specifically, when we have two feature functions $h_1 = logPr(F|E)$ and $h_2 = logPr(E)$, equation (1.2) is essentially equivalent to the equation (1.1), which means log-linear approach is a more general model. Besides the two common features in source-channel approach, i.e. language model and translation model, other possible features as suggested by (Och & Ney, 2002) are sentence length feature, additional language models (class-based five-gram language model), lexicon co-occurrence, lexical feature, and grammar feature.

As summarized in Figure 1.1, the two current models adopting this approach are Model 6 (Och & Ney, 2003), and factored translation model (Koehn & Hoang, 2007). In (Och & Ney, 2003) work, they realize that HMM model makes well use the locality in the *source* language, where as IBM Model 4 makes use of locality in the *target* language. As such, they have come up with this Model 6 by combining HMM and Model 4 in a log-linear way, and claimed to yield better results than the HMM model as well as the 5 IBM models. In Model 6, the authors also propose an efficient greedy search algorithm as suggested by (Brown et al., 1993), called *pseudo-Viterbi alignment*, which is simple-model Viterbi plus improving iterations. The factored translation model was recently proposed with the ambition to consider different aspects of translations at different levels such as morphological, syntactic, or semantic levels. Together

with the open-source MOSES systems (Koehn, Federico, Shen, Bertoldi, Bojar, Callison-Burch, Cowan, Dyer, Hoang, Zens, Constantin, Moran, & Herbst, 2007a), MOSES provides the flexibility to incorporate different features, which they call factors, in a log-linear manner. A work in their system is now not only a token, but a vector of different factors such as surface for, lemma, part-of-speech, or morphological features. They have showed that the factored translation model achieves better translation performance, both in terms of automatic scores, as well as grammatical coherence.

In log-linear systems, the weights λ_i are estimated using MERT (Minimum Error Rate Training) method(Och, 2003).

$$\lambda^* = \arg\min_{\lambda} \mathrm{Err}(\mathrm{cand}(\lambda), \mathrm{ref})$$

(MERT to be further explained)

1.1.3 State-of-the-art SMT systems

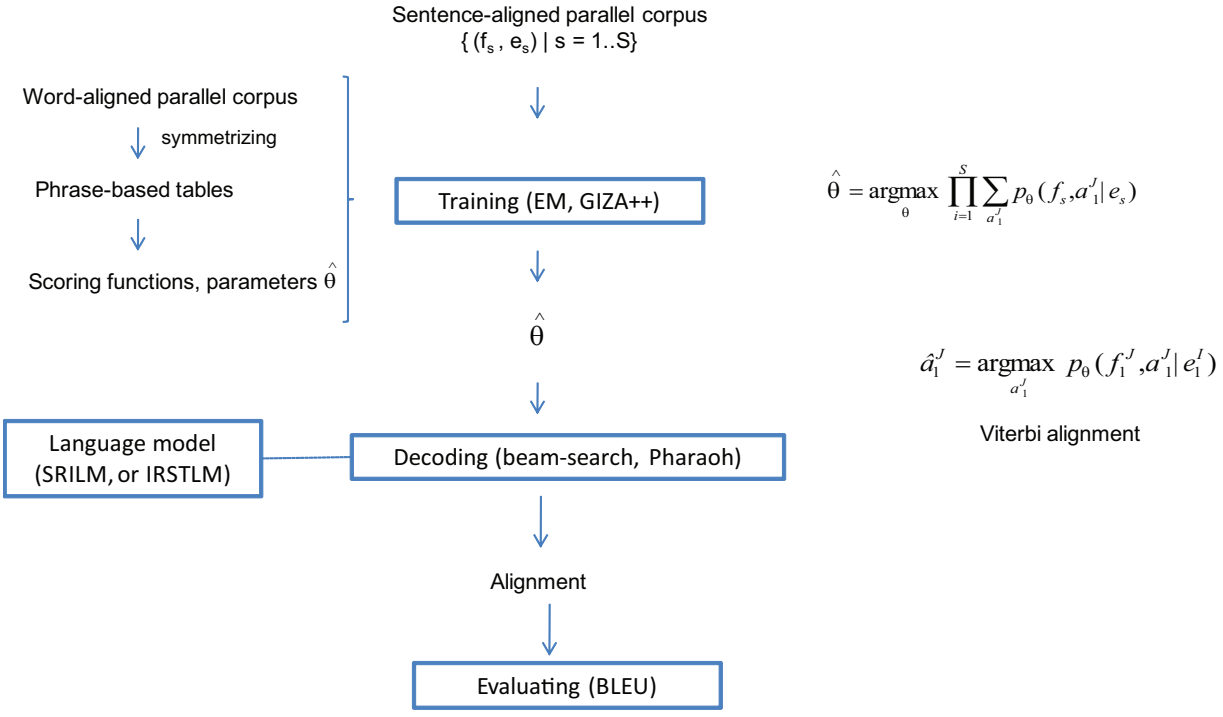

Figure 1.5: A sample SMT system with suggested publicly-available tools and methods

Evaluation

Good automatic scoring: Word Error Rate (WER), Position Independence Word Error Rate (PER), 100-BLEU score, NIST-score

BLEU (BiLingual Evaluation Understudy): measure the matches of short phrases between translated and reference text, as well as the difference in length of the reference and output.

Using unigrams, bigrams, trigrams, and 4-grams as well as penalty for too-short sentences to judge the precision of the system with respect to reference translations.

WER (Word Error Rate): measure the number of matching output and reference words with word order preserving, and maximum attainable number of single-word matches.

Minimum edit distance from the candidate translation to the reference translation. For mWER (multireference WER), multiple reference translations are used.

PER (Position Independence Word Error Rate): compare words in the two sentences, ignoring word order.

Smoothing

To overcome the problem of overfitting on the training data, and to enable the models to cope better with rare words, smooth the alignment and fertility probabilities.

Alignment probabilities are interpolated with uniform distribution p(i | j, I) = 1/I

Fertility probabilities are smoothed so that for rare words, the length of the words are taken into account.

The state-of-the-art smoothing technique is modified Kneser-Ney interpolation (Chen & Goodman, 1996)

Efficient method for growing and pruning Kneser-Ney smoothed models are presented in (Siivola, Hirsimaki, & Virpioja, 2007)

Training:

Parameters θ varies according to the system model, e.g, θ in model 4 consists of lexicon, alignment, and fertility parameters

EM algorithm: hidden variable is alignment. Use EM iteratively to estimate the model parameters, compute alignments from the estimated parameters, and use the alignments to re-estimate the parameters.

Viterbi alignment:

** Model 1 & 2: O(I.J)

** HMM: O(I^2, J)

** Model 3, 4, 5, 6: NP-complete

1.2 Morphological analysis in statistical machine translation

Figure 1.6: Language classification on morphological influence (Dyer, 2007a)

We begin this section with the classification of languages based on morphological influence adopted from (Dyer, 2007a) (see Figure 1.6). As we could observer languages vary in morphological degree from *isolating* (low-inflected) languages in which each word generally has one morpheme, like Vietnamese or Chinese, to *polysynthetic* (highly-inflected) languages where a single word could have many morphemes, like Siberian Yupik ("Eskimo") or Navaho. One of the biggest challenges for SMT systems in the morphology perspective is the problem of sparse data when dealing with highly-inflected languages. As for highly-synthetic languages, more and more morphemes could be added to a word to form a new word with more enriching the meaning, so many words tends to have frequency of one in the training corpus, or even does not appear at all, causing troubles for many SMT systems. Moreover, when translating from a low-inflected language to a highly-inflected language and vice versa, the problem of non-correspondence between words in two languages become more severe as a word in the low-inflected might corresponding to a suffix or prefix morpheme, or no correspondence at all.

These challenges suggest why there is a need to incorporate morphological knowledge in SMT systems.

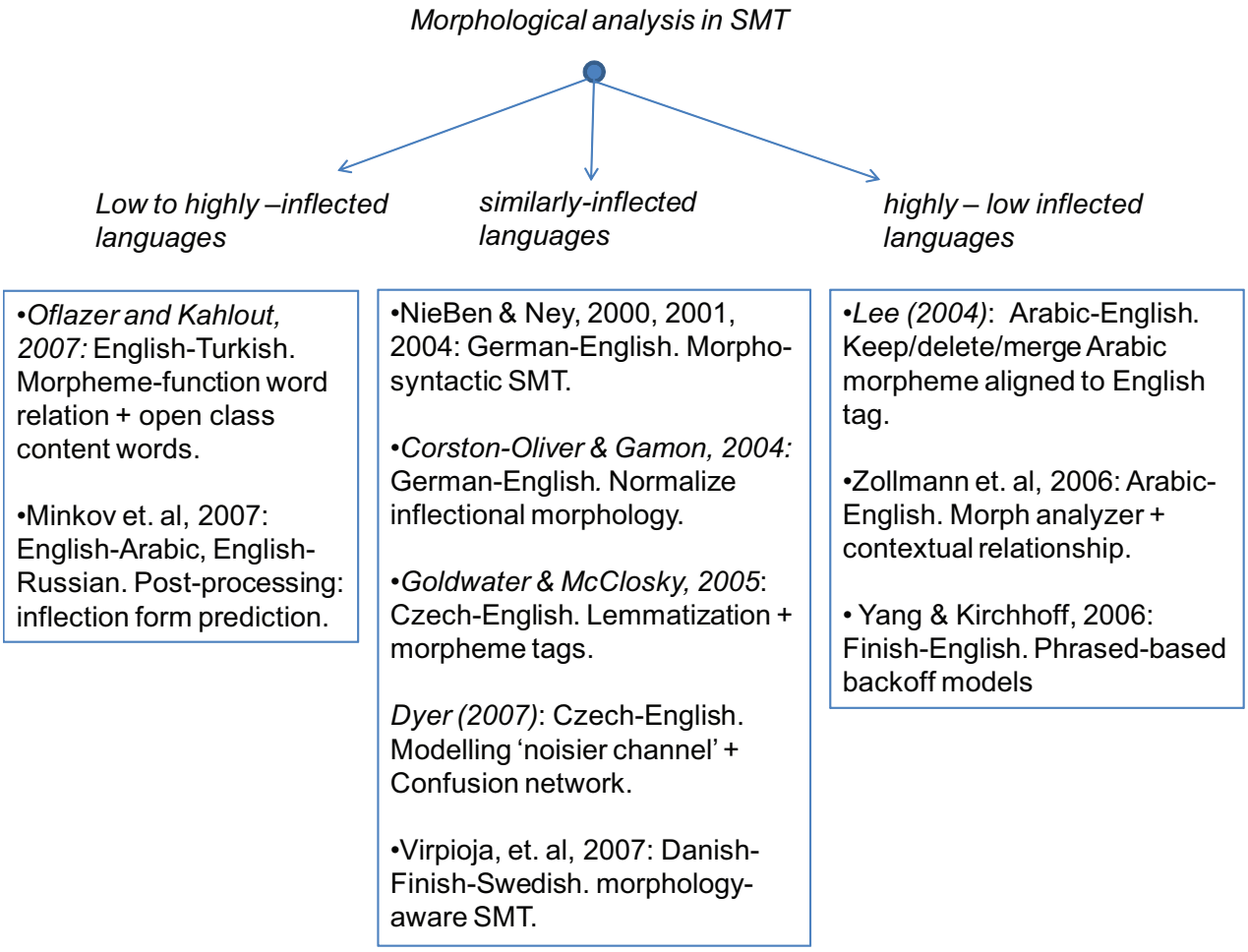

Figure 1.7: Morphological analysis in SMT

We analyze current morphological SMT systems based on their choices of languages either from morphologically poor language to morphologically rich language, vice versa, or from and to languages of similar morphological degree. We classify a pair of languages as *similarly-inflected* when they are less than two columns apart in the classification by Dyer (see Figure 1.6). In our figure 1.7 which summarizes different works on morphological SMT, readers could notice that many SMT systems have been constructed for those similarly-inflected language pairs such as German-English, or Czech-English. In (Nießen & Ney, 2000), (Nießen & Ney, 2001), (Nießen & Ney, 2004), morphological analysis is used to address the issue of restructuring in SMT, such as

question inversion or separated verb prefixes, which helps improving the translation performance from German to English. The authors have also suggested using hierarchical lexicon models to deal with the problem of data sparseness, resulting similar level of alignment quality with smaller corpora. In (Corston-Oliver & Gamon, 2004), normalization of inflectional morphology is experimented on different scenarios (e.g. stemming on verb, or noun phrase etc.). Even though shown to improve the perplexity of the models as well as reduces alignment errors when translating German-English, the system only translate into the base form, and not being able to decide the contextually appropriate word form. (Goldwater & McClosky, 2005) investigates in how morphological analysis alone could help MT system, and try to make Czech input date more English-like by suppressing unnecessary morphological distinctions and expressing necessary distinctions in ways that are similar to English. (Dyer, 2007b) suggests the use of morphological transformation under the view of a 'noisier channel', which extend the usual noisy channel by adding a morphological component to the channel pipeline. The morphological component allow the system to deal with ambiguous input in the form of *Confusion Network*. Based on a hierarchical phrase-based decoder, the system claims to obtain a significant BLEU score improvement when translating from Czech to English. Recently, the work (Virpioja, Vyrynen, Creutz, & Sadeniemi, 2007) has attempted to translate among highly-inflected Nordic languages (Danish, Finish, and Swedish) at morpheme level. Danish and Swedish are very close to each other in terms of grammar and vocabulary. Finish, on the other hand, is considerably different from Danish and Swedish, famous for its extremely rich morphology, and the most difficult language to translate from and to among those languages available in the Europarl corpus (Koehn, 2005). Even though their system did not obtain higher BLEU scores compared to the word-based approach, they have presented a promising unsupervised, language-dependent approach with the use of the unsupervised morphological analysis algorithm Morfessor (Creutz & Lagus, 2005), and variable n-gram model VariKN (Siivola et al., 2007).

The second class of works focus on translating from synthetic languages to isolating languages, which is considered a harder task than translating among similarly-inflected languages. This is due to the morphological complexity of the source language that is markedly different

from the target language, resulting in the un-correspondence problem in grammar and vocabulary, e.g. a word in Arabic may correspond to multiple words in English. Moreover, scarce resource is often a challenge when working with synthetic languages as not many languages have large training corpora. Even if corpora are available for a highly-inflected language, they may not be large enough to cover the majority of distinct words in the language due to its morphology abundance, which results in large vocabulary. We present here several recent works on the aforementioned direction. In (Lee, 2004), the authors presents a technique to induce a morphological and syntactic symmetry between Arabic and English, which presupposes a POS-tagged parallel corpus as well as pre-segmentation of Arabic words into prefix(es)-stem-suffix(es) form. By considering the consistency of an aligned English POS tag and an Arabic morpheme, the system determines whether that morpheme is to be kept, merged back to the original stem, or discarded. In (Zollmann, Venugopal, & Vogel, 2006), the task of translating Arabic to English is assisted by using Buckwalter Arabic Morphological Analyzer (BAMA)(Buckwalter, 2004). Based on the contextual relationship, e.g. word occurrences, in the target language the system determines the most appropriate segmentation of an Arabic word from those possible segmentations generated by BAMA tool. In (Yang & Kirchhoff, 2006), a phrase-based backoff model which performs morphological analysis is used to handle unseen word form in the source languages when translating from German to English, and Finish to English. Specifically, stemming and compound splitting operations are interleaved to hierarchically handle an unseen word. For stemming, the system makes uses of TreeTagger (Schmid, 1994) for German and the Snowball stemmer [2] for Finish. Compound splitting is accomplished by a simple technique of considering all possible segmenting ways and constraining on the lengths of each subpart.

So far, only related works on the two directions (among similarly-inflected languages, and low to highly-inflected languages) are mentioned, how about the last directions? In fact, very limited works on translating from isolating to synthetic language are available in the literature. Translating from information-poor into an information-rich language is inherently more difficult than the reverse direction as supported by Koehn in his comprehensive analysis on 110 SMT

[2] http://snowball.tartarus.org/

systems (Koehn, 2005). According to him, researchers have made a similar observation that Chinese-English SMT systems perform worse than Arabic-English SMT systems. In our belief, one of the main reason is that reducing redundancy in the morphologically rich source language before translating is easier than introducing additional information or combining words in the morphologically poor source language. By the time of this review, only two works on this direction are available, and have shown promising results. The first work (Oflazer & Durgar El-Kahlout, 2007) explores the translation task from English to Turkish based on an observation that a complete English phrase needs to be used to align with a Turkish word, and might be discontinuous on the English side. The main idea is to get open-class words in English aligned with stems in Turkish open-class words by separating additional "noise" from morphemes and other function words. TreeTagger (Schmid, 1994) is used to provide *lemma* and *part-of-speech* for each English word. For Turkish, their own morphological analyzers output several lexical morpheme segmentations for each word, which are then disambiguated by a external statistical disambiguator specially designed for Turkish. Additional data containing only open-class English words and Turkish open-class stems is augmented to the normal training corpus, which results good BLEU points for the system. The system uses morpheme-based language model, and rescore the outputs with word-based language model. In (Minkov, Toutanova, & Suzuki, 2007), the authors work on translation task from English to Arabic and Russian in which they have presented an interesting approach in predicting the inflected word forms of the target morphologically rich languages. The system preassumes available lexicons for both source and target languages that provides morphological information for words in source and target languages. The system aims at taking an output sentence from an MT, convert into stemmed-version sentence, and predict the correct inflected-version sentence by employing all information from lexical. That aim is accomplished by using a second-order probabilistic model decomposing overall probability to individual word predictions, as well as categorizing different features for probability prediction such as *monolingual* for target language, or bilingual for both languages.

Conclusion: to be done.

Chapter 2

Proposed systems

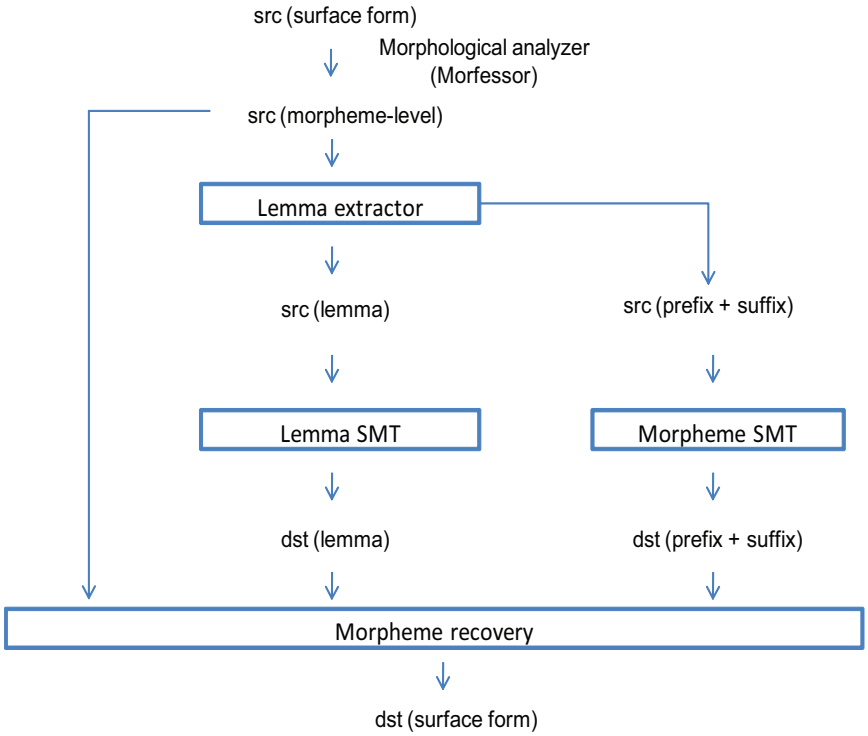

Figure 2.1: Translating from src to dst languages of similar inflection levels (low-low, high-high)

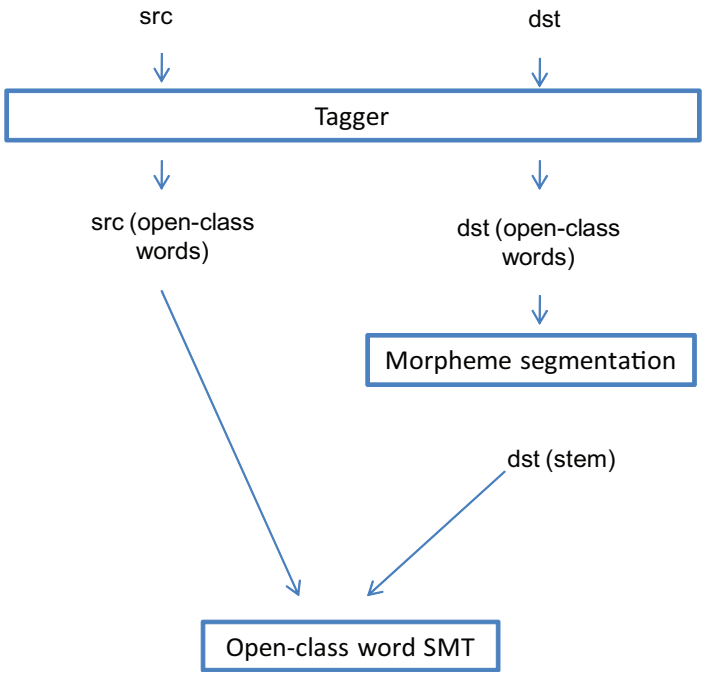

Figure 2.2: Training for src (lowly inflected) and dst (highly inflected)

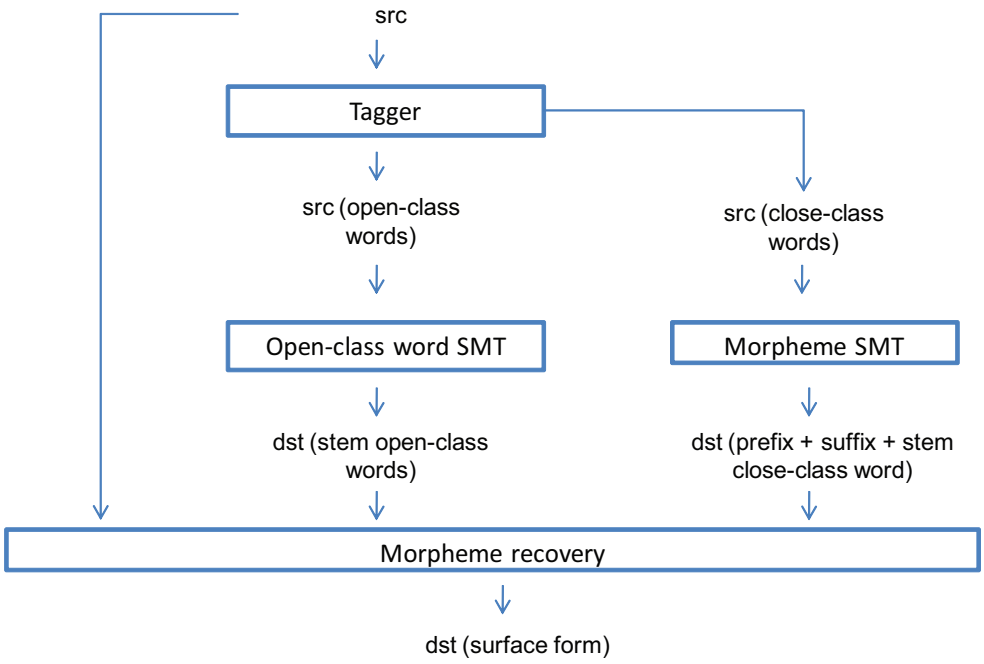

Figure 2.3: Translating from src (lowly inflected) to dst (highly inflected)

References

Brown, P. F., Pietra, S. D., Pietra, V. J. D., & Mercer, R. L. (1993). The mathematic of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*(2), 1993, 263–311.

Buckwalter, T. (2004). Buckwalter arabic morphological analyzer version 2.0. Linguistic Data Consortium, Philadelphia.

Chen, S. F., & Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In A. Joshi, & M. Palmer (Eds.), *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics* (pp. 310–318), San Francisco, 1996: Morgan Kaufmann Publishers.

Corston-Oliver, S., & Gamon, M. (2004). Normalizing german and english inflectional morphology to improve statistical word alignment. *AMTA* (pp. 48–57), 2004.

Creutz, M., & Lagus, K. (2005). Inducing the morphological lexicon of a natural language from unannotated text. , 2005.

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

Dyer, C. (2007a). Decoder guided backoff: using word lattices to improve translation from morphologically complex languages. Presented at the MT Marathon, Edinburgh University. http://www.ling.umd.edu/r̃edpony/edinburgh.pdf.

Dyer, C. J. (2007b). The "noisier channel": Translation from morphologically complex languages. *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 207–211), Prague, Czech Republic, June, 2007: Association for Computational Linguistics.

Goldwater, S., & McClosky, D. (2005). Improving statistical mt through morphological analysis. *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 676–683), Morristown, NJ, USA, 2005: Association for Computational Linguistics.

Jurafsky, D., & Martin, J. H. (2007). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, Chap. 25. Draft at 31 December, 2007. Available at http://www.cs.colorado.edu/m̃artin/slp2.html.

Knight, K. (1999). A statistical mt tutorial workbook.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Machine Translation Summit X* (pp. 79–86), Phuket, Thailand, 2005.

Koehn, P., Federico, M., Shen, W., Bertoldi, N., Bojar, O., Callison-Burch, C., Cowan, B., Dyer, C., Hoang, H., Zens, R., Constantin, A., Moran, C. C., & Herbst, E. (2007a). *Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding* (Technical report). Johns Hopkins University.

Koehn, P., & Hoang, H. (2007). Factored translation models. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 868–876), 2007.

Koehn, P., Hoang, H., Mayne, A. B., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007b). Moses: Open source toolkit for statistical machine translation. *Annual Meeting of the Association for Computation Linguistics (ACL), Demonstration Session* (pp. 177–180), Jun, 2007.

Lee, Y.-S. (2004). Morphological analysis for statistical machine translation. *In Proc. of NAACL*, Boston, MA, 2004.

Minkov, E., Toutanova, K., & Suzuki, H. (2007). Generating complex morphology for machine translation. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 128–135), Prague, Czech Republic, June, 2007: Association for Computational Linguistics.

Nießen, S., & Ney, H. (2000). Improving smt quality with morpho-syntactic analysis. *Proceedings of the 18th conference on Computational linguistics* (pp. 1081–1085), Morristown, NJ, USA, 2000: Association for Computational Linguistics.

Nießen, S., & Ney, H. (2001). Morpho-syntactic analysis for reordering in statistical machine translation. *MT Summit VIII* (pp. 247–252), Santiago de Compostela, Spain, September, 2001.

Nießen, S., & Ney, H. (2004). Statistical machine translation with scarce resources using morpho-syntactic information. *Comput. Linguist.*, *30*(2), 2004, 181–204.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* (pp. 160–167), Morristown, NJ, USA, 2003: Association for Computational Linguistics.

Och, F. J., & Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. *ACL 2002: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 295–302), Association for Computational Linguistics, Philadelphia, PA, July, 2002.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 2003, 19–51.

Oflazer, K., & Durgar El-Kahlout, I. (2007). Exploring different representational units in English-to-Turkish statistical machine translation. *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 25–32), Prague, Czech Republic, June, 2007: Association for Computational Linguistics.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *International Conference on New Methods in Language Processing*, Manchester, UK, 1994: unknown.

Siivola, V., Hirsimaki, T., & Virpioja, S. (2007). On growing and pruning kneserney smoothed n-gram models. *Audio, Speech, and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions*, Vol. 15 (pp. 1617–1274), July, 2007.

Virpioja, S., Vyrynen, J. J., Creutz, M., & Sadeniemi, M. (2007). Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. *Proceedings of the Machine Translation Summit XI* (pp. 491–498), Copenhagen, DenMark, 2007.

Vogel, S., Ney, H., & Tillmann, C. (1996). Hmm-based word alignment in statistical translation. *Proceedings of the 16th conference on Computational linguistics* (pp. 836–841), Morristown, NJ, USA, 1996: Association for Computational Linguistics.

Yang, M., & Kirchhoff, K. (2006). Phrase-based backoff models for machine translation of highly inflected languages. *EACL*, 2006.

Zollmann, A., Venugopal, A., & Vogel, S. (2006). Bridging the inflectional morphology gap for arabic statistical machine translation. *Proceedings of the HLT-NAACL* (pp. 201–204), New York City, June, 2006.