ChunkMT: Statistical Machine Translation with Richer Linguistic Knowledge

Philipp Koehn
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292
koehn@isi.edu

Kevin Knight
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292
knight@isi.edu

Abstract

Word based statistical machine translation has emerged as a robust method for building open domain machine translation systems. We point out some problems with the approach and propose to address them by including richer linguistic information: part of speech tags and syntactic chunks. We propose a model with separate components for sentence level reordering, chunk mapping, word translation, and exact phrase lookup. We report improved performance over IBM Model 4 on a short sentence translation task.

1 Motivation

1.1 Background

Research in statistical machine translation was pioneered by the Candide project at IBM [Brown et al., 1990]. In recent years, these methods were re-implemented, and software has become freely available in form of the training system Giza [Al-Onaizan et al., 1999] and decoders that perform actual translations [Germann et al., 2001]. Given these tools and a parallel corpus as training material, a statistical machine translation system can now be built within a few hours of computing time.

In this paper, we build upon ideas from the model proposed by IBM. Roughly speaking, the IBM model decomposes the machine translation process from a source language (say, German) to a target language (say, English) into the separate translation of single words (by a *word translation table*), with the possibility of German words being dropped (*null generated*) and English words added (*zero fertility words*). Multiple German words may also map to a single English word (based on the *fertility* of the English word), but not vice versa. Words may be moved around during translation (*distortion*). All these decisions are weighted by probabilities. An English language model (usually using word trigrams) helps in ensuring fluent output.

While being a somewhat crude method, word based statistical machine translation is very robust: Given proper parameter values, any translation from a German sentence to an English sentence is possible. Also, when mistakes are made, the system usually decays gracefully by translating other parts of the sentence correctly.

However, a number of translation phenomena pose serious challenges (see Figure 1): The first problem is with multiple English words being translated from a single German word, which is not allowed by the IBM alignment scheme. Translations of multiple word phrases which do not decompose easily into word for word translations because of non-compositional semantics is a second problem, as this is also not allowed by the IBM alignment scheme. Finally, a practical problem in the estimation of the parameters of the IBM model is that only reorderings local to an area of a few words can be estimated with any accuracy, making larger syntactic transformations difficult to capture.

It is very hard to address these challenges within the word based statistical machine translation framework: all of the parameters are tied to words, and these problems are tied to the behavior of groups of words which we will call chunks, and which might be described using linguistic vocabulary such as *verb group* or *subject*. The behavior of chunks is above the word level, and the IBM model fails to capture this behavior.

1.2 Additional Linguistic Knowledge

In this paper, we propose to integrate two linguistic concepts into the statistical machine

Multiple English words for one German word

German:	`Zeitmangel erschwert das Problem .`
Gloss:	LACK OF TIME MAKES MORE DIFFICULT THE PROBLEM .
Correct translation:	Lack of time makes the problem more difficult.
MT output:	`Time makes the problem .`

Phrasal translation

German:	`Eine Diskussion erübrigt sich demnach .`
Gloss:	A DISCUSSION IS MADE UNNECESSARY ITSELF THEREFORE .
Correct translation:	Therefore, there is no point in a discussion.
MT output:	`A debate turned therefore .`

Syntactic transformations

German:	`Das ist der Sache nicht angemessen .`
Gloss:	THAT IS THE MATTER NOT APPROPRIATE .
Correct translation:	That is not appropriate for this matter .
MT output:	`That is the thing is not appropriate .`
German:	`Den Vorschlag lehnt die Kommission ab .`
Gloss:	THE PROPOSAL REJECTS THE COMMISSION OFF .
Correct translation:	The commission rejects the proposal .
MT output:	`The proposal rejects the commission .`

Figure 1: Problematic Cases for Word Based Statistical Machine Translation

translation framework: part of speech (POS) tags and chunks. These concepts enable us to better address the hard cases in Figure 1.

Both concepts are well established in the computational linguistics community. Tools for tagging words with their corresponding POS (such as VVFIN for an finite verb) are widely available.

The objective of chunking is the detection of simple non-recursive verb, noun, prepositional, or other phrases. First introduced by Abney [1991], chunking is frequently used in information extraction. A recent competition on a common data set at CoNLL [Sang and Buchholz, 2000] brought together a wide range of tools for this task.

One can easily see how the problematic cases from Figure 1 can be described in terms of chunk transformation: For instance, the third example requires that the noun chunk `der Sache` is transformed into the prepositional chunk `for this matter` and moved to the end of the sentence.

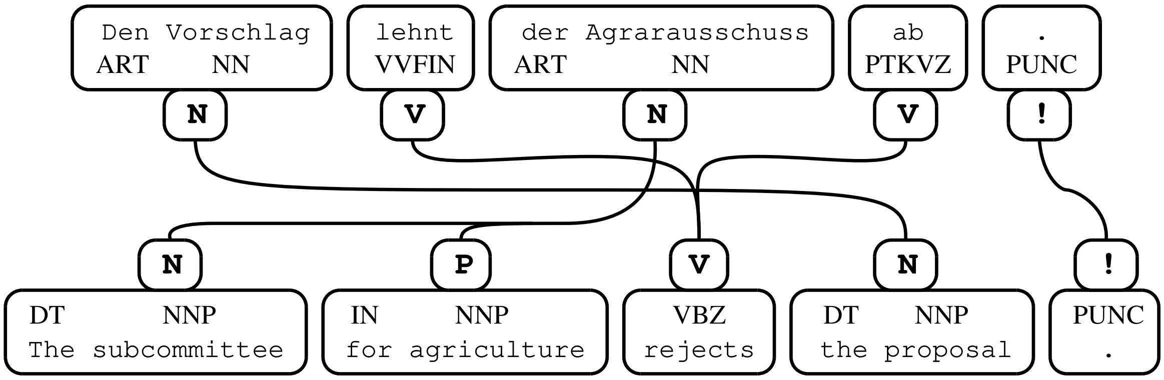
One reason we shy away from full parsing is that high quality parsers are not available for many languages. Another reason is that the transformation of parse trees during translation is a hard problem. Solutions to this were proposed by Wu [1997], Alshawi et al. [2000], and Yamada and Knight [2001]. One deficiency of their proposed solutions is that they only allow for reordering of children in the parse tree. Therefore, larger structural transformations (such as required by the second example in Figure 1) are out of reach. Our proposed approach could be altered to make use of full parsing, but for now we do without.

2 Overview of the Model

We decompose the task of machine translation into three steps (see Figure 2):

- sentence level chunk reordering
- chunk mapping
- word translation

The first component of our model – **sentence level chunk reordering** – takes as input a German sentence with an associated German chunk labeling. It then generates an English chunk label sequence and connects the English chunks to German chunks. Connections between the German and English chunks are



1 Sentence Level Chunk Reordering

This defines how chunks are connected from German to English language. Connections are many-to-many with transitive closure. Every chunk is connected.
This example has six different connections between a German and English chunk. They group into four **complex chunk** mappings.

	N	V	N	V	!
N			•		
P			•		
V		•		•	
N	•				
!					•

2 Chunk Mapping

Each complex chunk mapping defines an alignment of German to English POS tags.
The word translation below each chunk mapping matrix are given for clarification. Word choice takes place only in the last step.

N→N	ART	NN
DT	•	
NNP		•

den$_{\text{ART}}$ Vorschlag$_{\text{NN}}$
→ the$_{\text{DT}}$ proposal$_{\text{NNP}}$

V+V→V	VVFIN	PTKVZ
VBZ	•	•

lehnt$_{\text{VVFIN}}$ ab$_{\text{PTKVZ}}$ → rejects$_{\text{VBZ}}$

N→N+P	ART	NN
DT	•	
NNP		•
IN		
NNP		•

der$_{\text{ART}}$ Agrarausschuss$_{\text{NN}}$
→ the$_{\text{DT}}$ subcommittee$_{\text{NNP}}$ for$_{\text{IN}}$ agriculture$_{\text{NNP}}$

!→!	PUNC
PUNC	•

.$_{\text{PUNC}}$ → .$_{\text{PUNC}}$

3 Word Translations

Finally, word translations set the lexical composition of the English sentence. The English POS is already selected by the chunk mapping.

der	→	DT	the
Agrarausschuss	→	NNP	subcommittee
NULL	→	IN	for
Agrarausschuss	→	NNP	agriculture
lehnt, ab	→	VBZ	rejects
den	→	DT	the
vorschlag	→	NNP	proposal

Figure 2: The Three Steps of ChunkMT: Sentence Level Reordering, Chunk Mapping, and Word Translation. Exact phrase lookup can also be used in place of the chunk mapping and word translation steps.

many-to-many, and all chunks must be connected to at least one chunk in the other language.

Chunk mapping is the mapping of a German chunk (say, `N`) into a English chunk (say, `N`). This is formulated in terms of POS tags. In the given example (see Figure 2), this means that the German POS tags ART NN are aligned to the English POS tags DT NNP.

Such a chunk mapping may map multiple chunks (or, a **complex chunk**) like `V+N` into a complex chunk like `V+P`. The single chunks combined in a complex chunk do not have to be adjacent, as seen in the example of the German verb chunks `V+V`.

The chunk mapping not only defines the which POS chunks are generated, but also their alignment. This alignment is many-to-many. It is possible for a POS tag in one language to not be aligned to anything, for example the English preposition IN is not aligned with anything in the third example in Figure 2.

Finally in the last step, the actual English words have to be produced by **word translation**. This is restricted by the English POS tags and alignments back to German POS, as defined by the selected chunk mapping.

There is also an alternative to the chunk mapping and word translation steps for complex phrases: the entire complex chunk may be translated by **exact phrase lookup**. For instance, if we know that we can translate the entire `N` chunk `der Agrarausschuss` into the `N+P` chunk `the subcommittee for agriculture`, we may use this directly instead. This enables the translation of idiomatic phrases that are not easily explained by the translation of the parts.

Finally, we employ an English trigram **language model** to produce fluent output.

Related Work

The idea of using chunk mappings instead of distortion and fertility for the local reordering of words is inspired by work by Och and Ney [2000] and Wang [1998].

They learn mappings of word groups without knowledge of chunks and POS tags. The role of POS tags is taken on by word classes that are automatically trained from bilingual corpora. The use of POS tags or word classes is not in contradiction: POS tags capture important linguistic properties of words, while word classes may capture important translation behavior. Combining both concepts may yield higher performance than either alone.

Both Wang and Och allow arbitrary boundaries for their *alignment templates*, while our chunk mapping are restricted by chunk boundaries.

An advantage of the chunk mapping approach is that we can clearly label what the German complex chunks (e.g. `V+N`) and the English complex chunks (e.g. `V+P`) are. This allows us to have a stronger grasp at the reordering of these entities at the sentence level, since we know something about their syntactic role. In other words, it allows improved reorderings of the chunk sequence, that are driven by syntactic transformations as exemplified in Figure 1.

3 Mathematical Formulation

In the mathematical formulation of our model we follow the noisy channel model employed in other statistical machine translation work. That means that instead of estimating $p(e|g)$ – the best translation e for an input sentence g – directly, we apply Bayes rule and maximize $p(g|e) \times p(e)$.

Intuitively, this splits the model into a translation part $p(g|e)$ and a language model $p(e)$. For the latter, we use a traditional trigram language model.

The translation part is decomposed into sentence level reordering (SLR), chunk mapping (CM) and word translations (W):

$$\begin{aligned} p(g|e) = \quad & p(SLR|e) \times \\ & \textstyle\prod_i p(CM_i|e, SLR) \times \\ & \textstyle\prod_j p(W_{ij}|CM_i, SLR, e) \end{aligned}$$

Since POS tagging and chunking is deterministic, e represents not only the English words, but also their POS and groupings into chunks.

The sentence level chunk reordering (SLR) and word reordering within chunks (CM) is done with templates in the form of the matrices in Figure 2. Word translation (W) is done by a word-by-word translation table.

Since direct estimation of these probabilities would lead to extreme sparse data problems, the three conditional probability distributions are simplified, so that

- $p(SLR)$ is conditioned only on the English chunk label sequence as a whole
- $p(CM_i)$ is conditioned only on the German and English chunk labels in this mapping, and the POS tags in the English complex chunk
- $p(W_{ij})$ is conditioned only on the aligned English word and its POS tag

Each word alignment in a chunk mapping is factored in with a word translation probability. This mapping may also include unaligned German and English words within the chunks. Unaligned German words are factored in with the probability $p(g_k|ZFERT, gpos_k)$. Unaligned English words are factored in with the probability $p(NULL|e_k, gpos_k)$.

Instead of decomposing the word translations, we can alternatively perform a direct phrase lookup

$$p(W_{i1}, ..., W_{in}|CM_i, SLR, e)$$

This is simplified to be conditioned only on the selected chunk mapping and the English words within in (e_i).

The combined model of decomposed word translation and phrase translation takes the form of a back-off model (with λ fixed at 0.5):

$$\begin{array}{lll} & \lambda & p(W_{i1}, ..., W_{in}|CM_i, e_i) \\ + & (1-\lambda) & \prod_j p(W_{ij}|CM_i, e_i) \end{array}$$

4 Parameter Estimation

We estimate the parameters for our model from a word-aligned parallel corpus. Knowledge of chunks and POS tags should be useful when word aligning a corpus. However, we currently simply use the word alignment provided by the Giza toolkit, which makes no use of this.

Giza produces only many-to-one alignments from German to English. Therefore, we use a trick proposed by Och and Ney [2000]. We create two word alignments with Giza: many-to-one, and one-to-many by reversing the translation direction. The intersection of these alignments has high precision with some loss in recall. To regain recall we align previously unaligned words in both English and German by looking for neighboring English and German words which are aligned to one another in one of the original alignments.

Both sides of the corpus are also POS tagged and chunked. We make use of the following tools:

- English POS tagger by Brill [1995][1]
- German POS tagger TnT by Brants [2000]
- English chunk parser CASS[2]
- German noun chunker LoPar by Schmidt and Schulte im Walde [2000][3]

The output of the chunk parsers requires some post-processing (e.g., LoPar detects only noun chunks).

Chunk mappings are collected from the parallel corpus as follows: If a German chunk and an English chunk contain a German word and a English word that are aligned to each other, we connect the two chunks. Chunks that contain no aligned words are attached to other chunks based on a few manual rules. We then perform a transitive closure on the chunk alignments: if chunk g_i is aligned with e_x, g_j is aligned with e_x, and chunk g_i is aligned with e_y, then also chunk g_j is considered aligned with e_y, even if they do not contain any words aligned to each other.

For this data, we can now collect statistics on word translations (including $p(g_k|ZFERT, gpos_k)$ and $p(NULL|e_k, gpos_k)$), complex chunk mappings, and sentence level reordering. From this we obtain conditional probability distributions by maximum likelihood estimation. Since the data for exact phrase lookup is highly noisy, we smooth these probabilities by adding 10 to the denominator and discarding any exact phrase pair that occurs less than 4 times.

5 Decoder

Translation (or *decoding*) takes place in two steps: First a **sentence level template** (SLT) for the sentence level chunk reordering is chosen. Secondly, the English translation is constructed a word at a time from left to right. This is repeated for the top n (in our experiments 20)

[1] available at `http://www.cs.jhu.edu/~brill`

[2] available at `http://www.research.att.com/~abney`

[3] available at `http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/LoPar-en.html`

SLT for the given German chunk sequence. Ultimately, the translation with the overall best score is selected as system output.

The construction of the English sentence can be implemented efficiently by a Viterbi search using dynamic programming. Chunk mapping templates are selected as needed. Then, the word slots are filled with use of the word-by-word translation table and the language model.

For each partial translation (or *hypothesis*), the following information has be maintained:

- last two words created (needed by the language model)
- current chunk mapping template, if not complete
- score so far
- back pointer to best path
- position of last chunk
- position of last word created within chunk

In our experiments, we restricted translations to contiguous complex chunks in English. The resulting complexity of the hypothesis space at any given position in the English sentence is $O(V^2C)$, with V being the vocabulary size, C the number of applicable chunk mapping templates. Complexity with respect to sentence length is linear.

Furthermore, we only consider the top 100 bigrams for V^2 and the top 100 chunk mappings, therefore limiting the size of the hypothesis space at any given sentence position to 10,000. These simplifications have virtually no impact on finding the best translation, as our experiments show. For a given sentence level template, the translation process finishes in fractions of a second on a 1 GHz Pentium 4 system running Linux.

6 Results

We evaluated the proposed ChunkMT statistical machine translation method on a German-English bilingual corpus extracted from the European Parliament proceedings[4]. This corpus contains 16.6 million words per language and 620,015 aligned sentences pairs. Of these, 359,672 sentence pairs were used for training, after discarding long sentences (more than 30 words) and reserving a test set.

From these sentences we collect:

- 337,120 distinct sentence level templates
- 485,705 distinct chunk mapping templates
- 38,551 distinct phrase translations with at least 4 occurrences
- 554,211 distinct lexicon entries with 40,703 distinct English words and 97,897 distinct German words
- 23,820 distinct ZFERT entries for estimating $p(g_k|ZFERT, gpos_k)$
- 48,385 distinct NULL generation entries for estimating $p(NULL|e_k, gpos_k)$

A weakness of the model presented here are sparse data problems for the estimation of the probability distribution for sentence level templates. This limits the applicability of this approach to short sentences. The experiments reported herein are therefore carried out on 7-token sentences (usually 6 words and punctuation). For 670 of 736 of these sentences at least one SLT can be found. We evaluate our system only on this part of the test corpus. For the other sentences, alternative methods could be used as backup, e.g., an IBM Model 4 decoder.

As a scoring metric for these 670 sentences we use IBM BLEU [Papinini et al., 2001], which is an average of 1 to 4-gram precision with a brevity penalty for short output. The English output is taken from the parallel corpus. In addition, we manually checked the first 100 sentence translations for correctness.

We evaluate the performance of the proposed system against a stack decoder using IBM Model 4 [Germann et al., 2001], a word based statistical machine translation method.

The base system ChunkMT achieves superior results over the Model 4 stack decoder, both according to the BLEU metric as well as in the number of correct sentence translations (see Figure 3).

Leaving out exact phrase translations (ChunkMT w/o Phrases), the performance decreases significantly: While the number of correct sentence translations remains high, the BLEU score decreases to roughly the level of Model 4.

Both ChunkMT systems produce more syntactically well-formed sentences than Model 4.

7 Error Analysis

Since we effectively decomposed translation into three components – sentence level reordering,

[4]This parallel corpus in 11 languages is available in various formats at *(link suppressed for blind review)*

	correct sentences	IBM BLEU	n-gram precision				brevity penalty
			uni	bi	tri	quad	
IBM Model 4 Stack	**32**	**0.183**	59.4%	28.2%	16.5%	9.7%	0.803
ChunkMT w/o Phrases	**35**	**0.182**	59.0%	27.7%	15.9%	9.8%	0.808
ChunkMT	**36**	**0.198**	59.0%	28.8%	17.1%	11.1%	0.829

Figure 3: Evaluation of ChunkMT against IBM Model 4

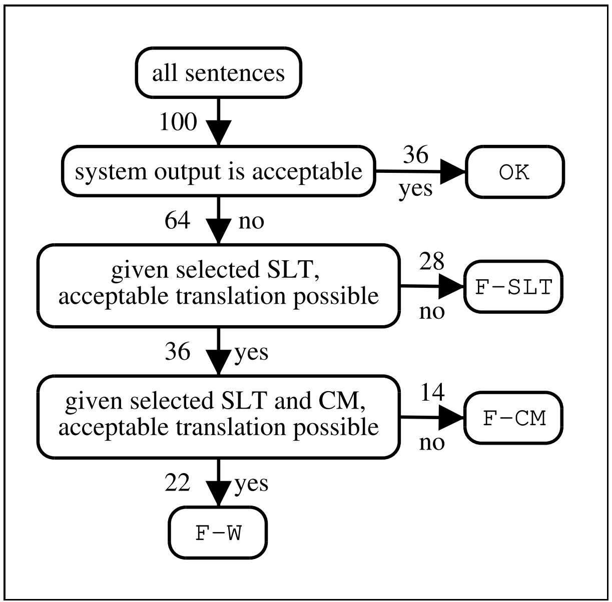


Figure 4: Error Analysis

chunk mapping, and word translation – we can attempt to pinpoint which of the components is the biggest source of error.

To speak of an error in an translation, one must have a clear idea of what correct output would look like. This is not a simple task, since often many acceptable translations are possible. In the extreme, one might say that it is impossible to translate a sentence at all, because some cultural connotation is always lost. Pragmatically, one might deem a translation acceptable if the majority of fluent speakers of both languages agrees. In the following analysis a translation is considered acceptable on the basis of the subjective judgment of the evaluator, who is fluent in both languages.

We classified the first 100 sentence translated by our system into four categories, described in Figure 4. First, we removed all acceptable translations from our analysis (category **OK**).

Then, we checked for each of the faulty translations if it is possible to manually create an acceptable translation within our model constraints under the condition that we use the same sentence level template. For instance, if the system elected to produce a sentence in the form `N V N !`, we checked if there is an English sentence consisting of a noun phrase, a verb phrase, another noun phrase, and punctuation, which is a acceptable translation and whose chunks are correspond to German chunks according to the mapping given by the sentence level template. If this is not possible, the failure is ascribed to the SLT selection (category **F-SLT**). The sentences in this category are also removed from further analysis.

The same is done for chunk mapping. For instance, if the chunk mapping requires an English noun chunk consisting of adjective and noun, we check if we can fill the POS slots with words that constitute a correct translation of the chunk and have plausible word level or phrase level alignments back to the German. Again, sentences for which there are chunk mappings where this is not possible are consider chunk mapping failures (**F-CM**) and removed from further analysis.

For the remaining sentences (**F-W**) only changes in word choice are needed to achieve an acceptable translation, since it is possible to construct a sentence with the given sentence level template and chunk mapping, and the given translation is not acceptable.

The biggest source of error even for these short sentences is the selection of an acceptable SLT (28 sentences). Note that for each SLT selection multiple CM selections and word choices have to be made, increasing the risk that any one of them will be false.

SLT errors are often due to sparse data: an acceptable SLT has not been seen in the training data, or has been seen very rarely. Other errors could be addressed by providing additional features. For instance, the necessary reordering for German object-verb-subject constructions is rarely done, because information about case markings of noun chunks is currently not used by the system.

Many of the sentences with chunk mapping errors (9 of the 14 sentences in the **F-CM** category) are caused by German words that translate into multiple English words. The translation of a German word into multiple English words is avoided, since the language model prefers short sentences. This could be addressed by a fertility feature as in the IBM Models. Also, it could be addressed by proper preprocessing: These fertile German words are often noun compounds such as `Massenkommunikation` that could be split up into `Masse Kommunikation`, which would be easier to translate into `mass communication`.

Many word translation errors are due to unknown German words (again, often noun compounds). But most errors are caused by multiple word senses and related issues such as anaphora resolution.

8 Conclusion

We proposed a new model for statistical machine translation and built a fast and high-accuracy decoder. Separate components – sentence level chunk reordering, chunk mapping, word translation, and exact phrase lookup – address different problems in the translation process. The goal of future research is to refine these components.

During sentence level chunk reordering, major syntactical transformations are carried out. As the proposed sentence level templates cause serious sparse data problems, we will have to decompose the sentence level templates into smaller steps for which richer statistics are available.

In future work, we will investigate further what drives these transformations, for instance whether there are different limitations to syntactic constructions in source and target language, or whether words that have no equivalent in the other language and require major reformulations.

By confining sentence level chunk reordering to the translation and reordering of chunk labels, we can turn it into a supervised learning problem with training and testing examples collected from a parallel corpus. Note that such a setup is not easily possible for the full translation task, since too many correct translations are possible, and it is too much to ask a system to come up with exactly the one provided in the parallel corpus.

The chunk mapping component addresses local word reorderings. It might also more easily deal with morphology than word based statistical machine translation [Nießen and Ney, 2001].

Finally, we feel that the translation of idiomatic or fixed expressions can best be handled by exact phrase lookup. It is hard to imagine how `kick the bucket` should be known to a system to translate as `den Löffel abgeben`, if this example has not been directly provided to it in some form. It is a challenging problem to learn these phrase translations, since the number of possible phrase alignments in a parallel corpus is huge. We showed that POS and chunking information is a useful filter.

References

Abney, S. (1991). Parsing by chunks. In *Robert Berwick, Steven Abney, and Carol Tenny: Principle-Based Parsing*. Kluwer Academic Publishers.

Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F.-J., Purdy, D., Smith, N. A., and Yarowsky, D. (1999). Statistical machine translation. Technical report, John Hopkins University Summer Workshop `http://www.clsp.jhu.edu/ws99/projects/mt/`.

Alshawi, H., Bangalore, S., and Douglas, S. (2000). Learning dependency translation models as collection of finite-state head transducers. *Computational Lingustics*, 26(1):45–60.

Brants, T. (2000). Tnt - a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP*.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Lingustics*, 21(4).

Brown, P., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Rossin, P. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):76–85.

Germann, U., Jahr, M., Knight, K., Marcu, D., and Yamada, K. (2001). Fast decoding and optimal decoding for machine translation. In *Proceedings of ACL 39*.

Nießen, S. and Ney, H. (2001). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. In *Workshop on Data-Driven Machine Translation at ACL 39*, pages 47–54.

Och, F. J. and Ney, H. (2000). Improved statistcal alignment models. In *Proceedings of ACL 38*.

Papinini, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). BLEU: a method for automatic evaluation of machine translation. Technical Report RC22176(W0109-022), IBM Research Report.

Sang, E. F. T. K. and Buchholz, S. (2000). Introduction to the CoNLL-2000 shared task: Chunking. In *Proceedings of CoNLL-2000 and LLL-2000*.

Schmidt, H. and Schulte im Walde, S. (2000). Robust German noun chunking with a probabilistic context-free grammar. In *Proceedings of COLING*.

Wang, Y.-Y. (1998). *Grammar Inference and Statistical Machine Translation*. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Lingustics*, 23(3).

Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of ACL 39*.