Statistical Machine Translation

Franz Josef Och and **Hermann Ney**
Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen - University of Technology
D-52056 Aachen, Germany
{och,ney}@informatik.rwth-aachen.de

Abstract

This paper gives an overview of statistical machine translation and presents the publically available SMT toolkit EGYPT. Starting with the Bayes decision rule as in speech recognition, we show how the required probability distributions can be structured into three parts: the language model, the alignment model and the lexicon model. We describe the components of the system and report results on the VERBMOBIL and the HANSARDS task. The experience obtained in the VERBMOBIL project, in particular a large-scale end-to-end evaluation, showed that the statistical approach resulted in significantly lower error rates than three competing translation approaches: the sentence error rate was 29% in comparison with 52% to 62% for the other translation approaches.

1 Introduction

Recently, statistical data analysis has been used to gather MT knowledge automatically, from parallel bilingual text. These techniques are extremely promising, as they provide a methodology for addressing the knowledge-acquisition bottleneck that plagues all large-scale natural language processing applications.

In the early 1990s, a substantial project by IBM achieved (and slightly exceeded) commercial-level translation quality through automatic bilingual-text analysis. Unfortunately, the statistical machine translation (SMT) techniques have not been applied widely in the MT community. This is partly due to the fact that the mathematics involved are not particularly familiar to computational linguistics researchers. Another reason is that common software tools and data sets are not generally available. It requires a great deal of work to build the necessary software infrastructure for experimentation in this area.

We will present an overview of the basic ideas in statistical machine translation, present recent promising results and describe the publically available SMT toolkit EGYPT.

2 Statistical Decision Theory and Linguistics

2.1 The Statistical Approach

The use of statistics in computational linguistics has been extremely controversial for more than three decades. The controversy is very well summarized by the statement of Chomsky in 1969 (Chomsky, 1969):

> "It must be recognized that the notion of a 'probability of a sentence' is an entirely useless one, under any interpretation of this term".

This statement was considered to be true by the majority of experts from artificial intelligence and computational linguistics, and the concept of statistics was banned from computational linguistics for many years.

What is overlooked in this statement is the fact that, in an automatic system for speech recognition or text translation, we are faced with the problem of taking decisions. It is exactly here where statistical decision theory comes in.

For the 'low-level' description of speech and image signals, it is widely accepted that the statistical framework allows an efficient coupling between the observations and the models, which is often described by the buzz word 'subsymbolic processing'. But there is another advantage in using probability distributions in that they offer an explicit formalism for expressing and combining hypothesis scores:

- The probabilities are directly used as scores: These scores are normalized, which is a desirable property: when increasing the score for a certain element in the set of all hypotheses, there must be one or several other elements whose scores are reduced at the same time.

- It is straightforward to combine scores: depending on the task, the probabilities are either multiplied or added.
- Weak and vague dependencies can be modeled easily. Especially in spoken and written natural language, there are nuances and shades that require 'grey levels' between 0 and 1.

Even if we think we can manage without statistics, we will need models which always have some free parameters. Then the question is how to train these free parameters. The obvious approach is to adjust these parameters in such a way that we get optimal results in terms of error rates or similar criteria on a representative sample. So we have made a complete cycle and have reached the starting point of the statistical modeling approach again.

When building an automatic system for speech or language, we should try to use as much prior knowledge as possible about the task under consideration. This knowledge is used to guide the modeling process and to enable improved generalization with respect to unseen data. Therefore in a good statistical modeling approach, we try to identify the common patterns underlying the observations, i.e. to capture dependencies between the data in order to avoid the pure 'black box' concept.

2.2 Bayes Decision Rule and System Architecture

In machine translation, the goal is the translation of a text given in a source language into a target language. We are given a source string $f_1^J = f_1 ... f_j ... f_J$, which is to be translated into a target string $e_1^I = e_1 ... e_i ... e_I$. Among all possible target strings, we will choose the string with the highest probability which is given by Bayes decision rule (Brown et al., 1993):

$$\begin{aligned} \hat{e}_1^I &= \arg\max_{e_1^I} \{Pr(e_1^I | f_1^J)\} \\ &= \arg\max_{e_1^I} \{Pr(e_1^I) \cdot Pr(f_1^J | e_1^I)\} \quad . \end{aligned}$$

Here, $Pr(e_1^I)$ is the language model of the target language, and $Pr(f_1^J | e_1^I)$ is the string translation model. The argmax operation denotes the search problem, i.e. the generation of the output sentence in the target language. The overall architecture of the statistical translation approach is summarized in Figure 1.

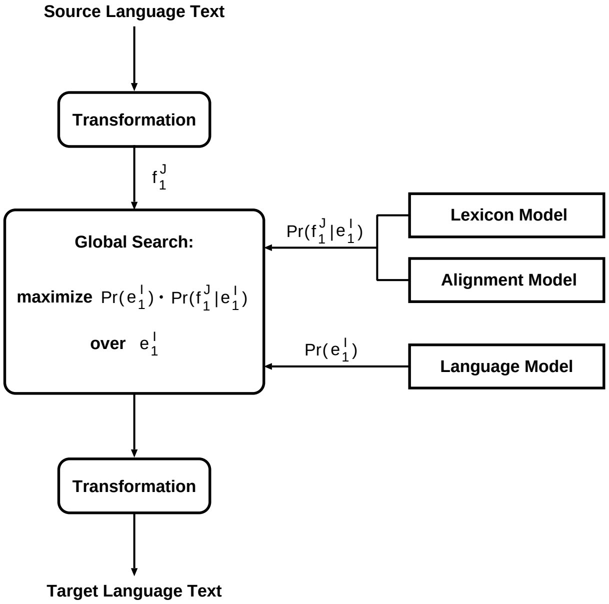

Figure 1: Architecture of the translation approach based on Bayes decision rule.

3 Alignment Modeling

3.1 Concept

A key issue in modeling the string translation probability $Pr(f_1^J | e_1^I)$ is the question of how we define the correspondence between the words of the target sentence and the words of the source sentence. In typical cases, we can assume a sort of pairwise dependence by considering all word pairs (f_j, e_i) for a given sentence pair $(f_1^J; e_1^I)$. Here, we will further constrain this model by assigning each source word to *exactly one* target word. Later, this requirement will be relaxed. Models describing these types of dependencies are referred to as *alignment models* (Brown et al., 1993; Vogel et al., 1996).

When aligning the words in parallel texts, we typically observe a strong localization effect. Figure 2 illustrates this effect for the language pair German–English. In many cases, although not always, there is an additional property: over large portions of the source string, the alignment is monotone.

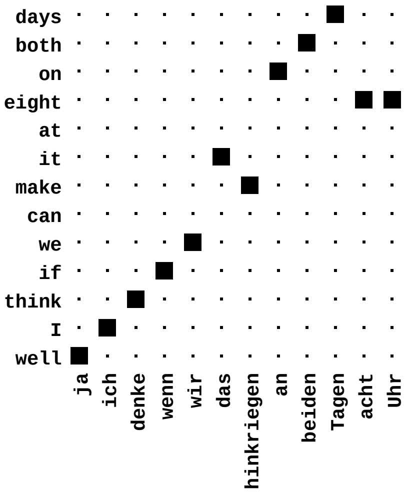

Figure 2: Word-to-word alignment.

3.2 Basic Models

To arrive at a quantitative specification, we define the alignment mapping: $j \rightarrow i = a_j$, which assigns a word f_j in position j to a word e_i in position $i = a_j$. We rewrite the probability for the translation model by introducing the 'hidden' alignments $a_1^J := a_1 ... a_j ... a_J$ for each sentence pair $(f_1^J; e_1^I)$. To structure this probability distribution, we factorize it over the positions in the source sentence and limit the alignment dependencies to a first-order dependence and arrive at the following model for $Pr(f_1^J | e_1^I)$:

$$p(J|I) \cdot \sum_{a_1^J} \prod_{j=1}^{J} [p(a_j | a_{j-1}, I, J) \cdot p(f_j | e_{a_j})] \ .$$

Here, we have the following probability distributions:

- the sentence length probability: $p(J|I)$, which is included here for completeness, but can be omitted without loss of performance;
- the lexicon probability: $p(f|e)$;
- the alignment probability: $p(a_j | a_{j-1}, I, J)$.

By making the alignment probability $p(a_j | a_{j-1}, I, J)$ dependent on the jump width $a_j - a_{j-1}$ instead of the absolute positions a_j, we obtain the so-called homogeneous hidden Markov model, for short HMM (Vogel et al., 1996).

In (Brown et al., 1993) were presented the models IBM-1 to IBM-5 which provide different decompositions of the probability $Pr(f_1^J, a_1^J | e_1^I)$: describing the probability of an alignment.

- In IBM-1 all alignments have the same probability.
- IBM-2 uses a zero-order alignment model $p(a_j | j, I, J)$ where different alignment positions are independent from each other.
- In IBM-3 we have an (inverted) zero-order alignment model $p(j | a_j, I, J)$ with an additional fertility model $p(\phi | e)$ which describes the number of words ϕ aligned to an English word e.
- In IBM-4 we have an (inverted) first-order alignment model $p(j | j')$ and a fertility model $p(\phi | e)$.
- The models IBM-3 and IBM-4 are deficient as they waste probability mass on non-strings. IBM-5 is a reformulation of IBM-4 with a suitably refined alignment model in order to avoid deficiency.

So the main differences of these models lie in the alignment model (which may be zero-order or first-order), in the existence of an explicit fertility model and whether the model is deficient or not.

3.3 Alignment Template Approach

A general shortcoming of the baseline alignment models is that they are mainly designed to model the lexicon dependences between single words. Therefore, we have extended the approach to handle word groups or phrases rather than single words as the basis for the alignment models (Och et al., 1999). In other words, a whole group of adjacent words in the source sentence may be aligned with a whole group of adjacent words in the target language. As a result, the context of words tends to be explicitly taken into account, and the differences in local word orders between source and target languages can be learned explicitly. Figure 3 shows some of the extracted alignment templates for a sentence pair from the VERBMOBIL training corpus.

4 Training

A main advantage of the statistical approach to machine translation lies in the fact that the

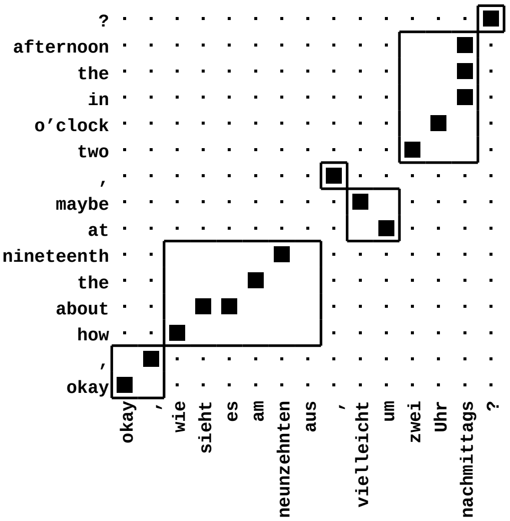

Figure 3: Example of a word alignment and of extracted alignment templates.

knowledge sources, e.g. translation model and language model, could be trained automatically by making use of a set of translation examples. The models described in the previous section contain a large set of free parameters. The training problem is an optimization problem to find the set of parameters which best explains the training data.

The training of all single-word based alignment models is done by the EM-algorithm using a parallel training corpus $(\mathbf{f}^{(s)}, \mathbf{e}^{(s)})$, $s = 1, \ldots, S$. In the E-step the counts for one sentence pair $(\mathbf{f}, \mathbf{e})$ are calculated. For the lexicon parameters the counts are:

$$c(f|e; \mathbf{f}, \mathbf{e}) \;=\; \sum_{\mathbf{a}} Pr(\mathbf{a}|\mathbf{f}, \mathbf{e}) \sum_{i,j} \delta(f, f_j)\delta(e, e_{a_j})$$

In the M-step the lexicon parameters are:

$$p(f|e) \;\propto\; \sum_{s} c(f|e; \mathbf{f}^{(s)}, \mathbf{e}^{(s)})$$

Correspondingly, the alignment and fertility probabilities can be estimated.

The models IBM-1, IBM-2 and HMM have a particularly simple mathematical form so that the EM algorithm can be performed exactly, i.e. in the E-step it is possible to efficiently consider all alignments. For the HMM we do this using the Baum-Welch algorithm (Baum, 1972). For IBM-3, IBM-4 and IBM-5 the count collection is performed only over a small number of good alignments.

The training algorithm for the alignment templates extracts all phrase pairs which are aligned in the training corpus up to a maximum length of 7 words. To improve the generalization capability of the alignment templates, the templates are determined for word classes rather than words directly. These word classes are determined by an automatic clustering procedure (Och, 1999). The training of the alignment templates is described in more details in (Och et al., 1999).

5 Search

The task of the search algorithm is to generate the most likely target sentence e_1^I of unknown length I for an observed source sentence f_1^J. The search must make use of all three knowledge sources as illustrated by Figure 4: the alignment model, the lexicon model and the language model. All three of them must contribute in the final decision about the words in the target language.

To illustrate the specific details of the search problem, we use the *inverted* alignment: $i \rightarrow j = b_i$, which is a mapping from *target* to *source* positions rather the other way round. We replace the sum over all alignments by the best alignment, which is referred to as maximum approximation in speech recognition. Using a bigram language model $p(e_i|e_{i-1})$, we obtain the following search criterion:

$$\max_{b_1^I, e_1^I} \prod_{i=1}^{I} [p(e_i|e_{i-1}) \cdot p(b_i|b_{i-1}, J) \cdot p(f_{b_i}|e_i)]$$

Considering this criterion, we can see that we can build up hypotheses of partial target sentences in a *bottom-to-top* strategy over the positions i of the target sentence e_1^i as illustrated in Figure 5. An important constraint for the alignment is that *all* positions of the source sentence should be covered exactly *once*. This constraint is similar to that of the traveling salesman problem where each city has to be visited exactly. It has been shown that the decoding problem is NP complete (Knight, 1999). Details on various search strategies can be found in (Ney et al., 2000).

6 The EGYPT toolkit

The SMT techniques have unfortunately not been applied widely in the MT community. This

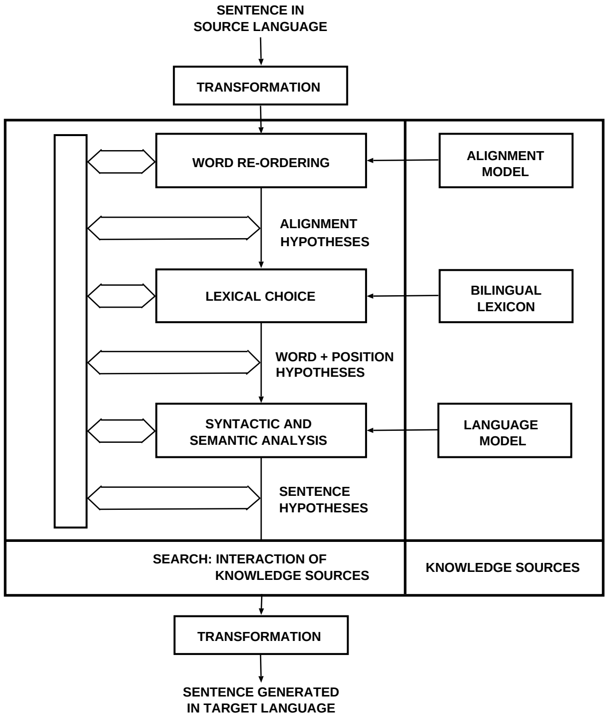
SENTENCE IN SOURCE LANGUAGE
TRANSFORMATION
WORD RE-ORDERING
ALIGNMENT MODEL
ALIGNMENT HYPOTHESES
LEXICAL CHOICE
BILINGUAL LEXICON
WORD + POSITION HYPOTHESES
SYNTACTIC AND SEMANTIC ANALYSIS
LANGUAGE MODEL
SENTENCE HYPOTHESES
SEARCH: INTERACTION OF KNOWLEDGE SOURCES
KNOWLEDGE SOURCES
TRANSFORMATION
SENTENCE GENERATED IN TARGET LANGUAGE

Figure 4: Illustration of search in statistical translation.

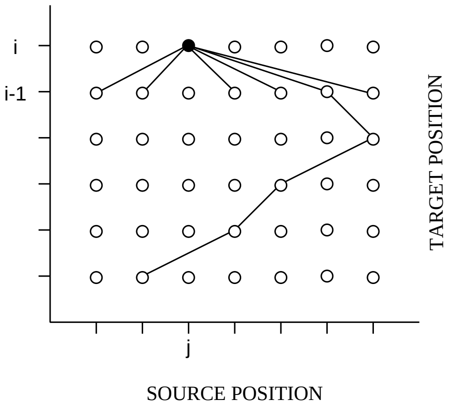

Figure 5: Illustration of bottom-to-top search.

is partly due to the fact that the mathematics involved are not particularly familiar to computational linguistics researchers. Another reason is common software tools and data sets are not generally available. It requires a great deal of work to build the necessary software infrastructure for experimentation in this area.

Such infrastructure is available now using the publically available toolkit EGYPT, which has been constructed in a six-week summer-workshop at Johns Hopkins University (Al-Onaizan et al., 1999)[1]. The main part of the software is the training program which includes training, data preparation, a sophisticated graphical interface for browsing word-by-word alignments and bilingual corpora and other tools. In the near future it will also include decoding software for performing actual translation.

In the following, we give a short description of the core software modules developed at the workshop:

Giza

The program GIZA is the training program for the alignment models described in section 3. It currently deals with the model IBM-1 to IBM-3. In the near future it will be available an extended version including the models IBM-4 and IBM-5, faster training and additional features like a correct implementation of 'pegging' (Och and Ney, 2000b).

Weaver

The program WEAVER is the decoding program. It is based on the stack decoding paradigm. It is planned to integrate WEAVER in future releases of the EGYPT toolkit.

Cairo

The program CAIRO is a visualization tool developed for word alignments. It helps to visualize alignments and the probability distributions occurring in IBM-3.

Whittle

The program WHITTLE is a corpus preparation tool. It splits the corpus into training and test corpora, generates vocabulary files and is able to write the corpus format that is needed by GIZA.

7 Experimental Results

We present results on the VERBMOBIL and the HANSARDS task (Table 2). For both tasks we manually aligned a randomly chosen subset of the training corpus (Table 1). From this corpus the first 100 sentences were used as validation corpus to optimize the smoothing parameters and the remaining sentences were used as test corpus.

[1] The toolkit could be downloaded from `http://www.clsp.jhu.edu/ws99/projects/mt/toolkit/`

Table 2: Training corpora sizes.

	Languages		Words		Vocabulary	
Corpus	SL/TL	Sentences	SL	TL	SL	TL
VERBMOBIL	English/German	34k	343 076	329 625	3 505	5 936
HANSARDS(50k)	French/English	50k	825 713	751 849	19 900	25 000
HANSARDS(200k)	French/English	200k	3 273 640	2 980 160	44 475	34 865
HANSARDS(500k)	French/English	500k	8 173 413	7 440 097	64 293	50 323
HANSARDS(1500k)	French/English	1500k	24 338 195	22 163 092	100 270	78 333

Table 1: Manually annotated test corpora.

	Words		
Corpus	SL	TL	Sentences
VERBMOBIL	3233	3109	354
HANSARDS	8749	7946	500

7.1 Alignment Quality

Evaluation Criterion

In the following, we present an annotation scheme for single-word based alignments and a corresponding evaluation criterion.

We developed an annotation scheme for word alignments that makes it possible to annotate explicitly the ambiguous alignments. We allowed human experts who performed the annotation to specify two different kinds of alignments: an S (sure) alignment which is used for alignments that are unambiguous and a P (possible) alignment which is used for alignments that might or might not exist. The P relation is used especially to align words within idiomatic expressions, free translations, and missing function words ($S \subseteq P$).

The thus obtained reference alignment may contain many-to-one and one-to-many relationships. Figure 6 shows an example of a manually aligned sentence with S and P relations.

The quality of an alignment $A = \{(j, a_j) | a_j > 0\}$ is then computed by appropriately redefined precision and recall measures:

$$recall = \frac{|A \cap S|}{|S|}, \quad precision = \frac{|A \cap P|}{|A|}$$

and the following error rate:

$$AER(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \quad .$$

Thereby, a recall error can only occur if a S(ure) alignment is not found and a precision error can only occur if a found alignment is not even P(ossible).

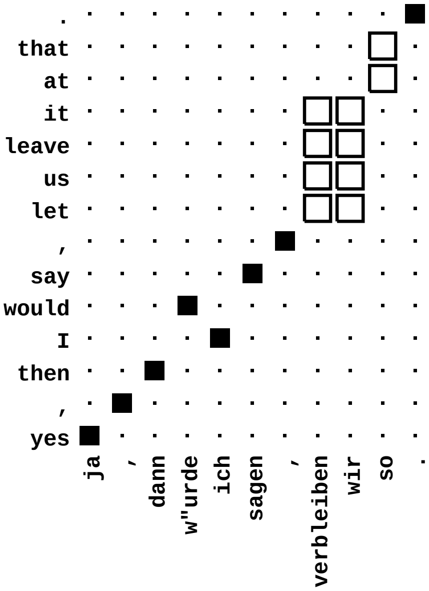

Figure 6: Example of a manual alignment with *S(ure)* (filled dots) and *P(ossible)* connections.

Table 3: Effect of using different amount of training data (HANSARDS task, training scheme 1 → HMM → 4).

	AER [%]		
Corpus	IBM-1	HMM	IBM-4
HANSARDS(50k)	34.3	18.0	15.6
HANSARDS(200k)	31.3	14.3	12.5
HANSARDS(500k)	30.3	12.8	10.7
HANSARDS(1500k)	29.4	11.0	9.4

More details to the evaluation methodology can be found in (Och and Ney, 2000a; Och and Ney, 2000b).

Results

Table 3 shows the effect of using different amounts of training data. As expected, more training data helps to improve alignment qual-

Table 4: Alignment quality in last iteration of IBM-4 of both translation directions.

	SL → TL			TL → SL		
Corpus	prec	rec	AER	prec	rec	AER
VERBMOBIL	93.2	95.5	5.8	90.0	87.9	10.9
HANSARDS(50k)	80.5	91.2	15.6	80.0	90.8	16.0
HANSARDS(200k)	84.3	93.1	12.5	84.2	93.4	12.4
HANSARDS(500k)	86.5	94.2	10.7	86.9	94.4	10.3
HANSARDS(1500k)	88.1	94.9	9.4	88.5	95.0	9.0

ity for all models. However, for IBM-1 the relative improvement is very small compared to the relative improvement using HMM and IBM-4. We conclude that more sophisticated alignment models are crucial for good alignment quality.

Looking at the AER obtained for both translation directions in Table 4 we see that for the language pair German-English (VERBMOBIL task) we observe that by using German as source language the AER is much higher than by using English as source language. This is because the baseline alignment representation as a vector a_1^J does in that case forbid that the often occurring German word compounds align to more than only one English word. Methods to deal with this asymmetry are described in (Och and Ney, 2000b).

7.2 Translation Quality

In order to show the performance of SMT in a real translation task we present here the results obtained of the Alignment Template system within the VERBMOBIL system which is a speech translation task in the domain of appointment scheduling, travel planning, and hotel reservation

This end-to-end evaluation of the VERBMOBIL system was performed at the University of Hamburg (Tessiore and v. Hahn, 2000).

Three other MT approaches had been integrated into the VERBMOBIL prototype system:

- a classical transfer approach (Becker et al., 2000; Emele et al., 2000; Uszkoreit et al., 2000),
  which is based on a manually designed analysis grammar, a set of transfer rules, and a generation grammar,
- a dialog-act based approach (Reithinger and Engel, 2000),
  which amounts to a sort of *slot filling* by classifying each sentence into one out of a small number of possible sentence patterns and filling in the slot values,
- an example-based approach (Auerswald, 2000),
  where a sort of nearest neighbor concept is applied to the set of bilingual training sentence pairs after suitable preprocessing.

In the final end-to-end evaluation human evaluators judged the translation quality for each of the four translation results using the following criterion:
Is the sentence approximatively correct: yes/no?
The evaluators were asked to pay particular attention to the semantic information (e.g. date and place of meeting, etc) contained in the translation. The evaluation was based on 9205 dialog turns. The speech recognizers used had a word error rate of about 25%. The sentence error rates are summarized in Table 5. As we can see, the error rates for the statistical approach are smaller by a factor of about 2 in comparison with the other approaches.

Table 5: Sentence error rates of end-to-end evaluation (speech recognizer with WER=25%).

Translation Method	Error [%]
Semantic Transfer	62
Dialog Act Based	60
Example Based	52
Statistical	29

In agreement with other evaluation experiments, these experiments show that the statistical modeling approach may be comparable to or better than the conventional rule-based approach. In particular, the statistical approach seems to have the advantage if robustness is important, e.g. when the input string is not grammatically correct or when it is corrupted by recognition errors.

8 Summary

In this paper, we have given an overview of the statistical approach to machine translation. We have presented various statistical alignment models of various complexity and described the basic concepts of training and search. We have given an overview of EGYPT, a publically available SMT toolkit. We have given results with respect to alignment and translation quality. The comparative evaluations with other translation approaches of the Verbmobil prototype system show that the statistical translation is superior, especially in the presence of speech input and ungrammatical input.

Acknowledgment

The work reported here has been supported partly by the Verbmobil project (contract number 01 IV 701 T4) by the German Federal Ministry of Education, Science, Research and Technology and as part of the EuTrans project (ESPRIT project number 30268) by the European Community.

References

Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, I. D. Melamed, F. J. Och, D. Purdy, N. A. Smith, and D. Yarowsky. 1999. Statistical machine translation, final report, JHU workshop. `http://www.clsp.jhu.edu/ws99/projects/mt/final_report/mt-final-report.ps`.

M. Auerswald. 2000. Example-based machine translation with templates. In Wahlster (Wahlster, 2000), pages 418–427.

L.E. Baum. 1972. An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes. *Inequalities*, 3:1–8.

T. Becker, A. Kilger, P. Lopez, and P. Poller. 2000. The Verbmobil generation component VM-GECO. In Wahlster (Wahlster, 2000), pages 481–496.

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

N. Chomsky. 1969. Quine's empirical assumptions. In D. Davidson and J. Hintikka, editors, *Words and objections. Essays on the work of W. V. Quine*. Reidel, Dordrecht, The Netherlands.

M. C. Emele, M. Dorna, A. Lüdeling, H. Zinsmeister, and C. Rohrer. 2000. Semantic-based transfer. In Wahlster (Wahlster, 2000), pages 359–376.

K. Knight. 1999. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4).

H. Ney, S. Nießen, F. J. Och, H. Sawaf, C. Tillmann, and S. Vogel. 2000. Algorithms for statistical translation of spoken language. *IEEE Trans. on Speech and Audio Processing*, 8(1):24–36, 1.

F. J. Och and H. Ney. 2000a. A comparison of alignment models for statistical machine translation. In *Proc. of the 18th Int. Conf. on Computational Linguistics*, pages 1086–1090, Saarbrücken, Germany, August.

F. J. Och and H. Ney. 2000b. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, pages 440–447, Hongkong, China, October.

F. J. Och, C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine translation. In *In Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, USA, June.

F. J. Och. 1999. An efficient method to determine bilingual word classes. In *EACL '99: Ninth Conf. of the Europ. Chapter of the Association for Computational Linguistics*, Bergen, Norway, June.

N. Reithinger and R. Engel. 2000. Robust content extraction for translation and dialog processing. In Wahlster (Wahlster, 2000), pages 428–437.

L. Tessiore and W. v. Hahn. 2000. Functional validation of a machine translation system: Verbmobil. In Wahlster (Wahlster, 2000), pages 611–631.

H. Uszkoreit, D. Flickinger, W. Kasper, and I. A. Sag. 2000. Deep linguistic analysis with HPSG. In Wahlster (Wahlster, 2000), pages 216–263.

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING '96: The 16th Int. Conf. on Computational Linguistics*, pages 836–841, Copenhagen, August.

W. Wahlster, editor. 2000. *Verbmobil: Foundations of speech-to-speech translations*. Springer-Verlag, Berlin.