

17th ICCRTS
“Operationalizing C2 Agility”

**Fast Realization of Automatic Translation Systems
for New Mission-Relevant Languages**

Topic 3: Data, Information and Knowledge
Topic 7: Architectures, Technologies, and Tools

Dr. Matthias Hecking (POC)

Fraunhofer Institute for Communication,
Information Processing and Ergonomics FKIE
Neuenahrer Straße 20, 53343 Wachtberg, Germany
matthias.hecking@fkie.fraunhofer.de

Sandra Noubours

Fraunhofer Institute for Communication,
Information Processing and Ergonomics FKIE
Neuenahrer Straße 20, 53343 Wachtberg, Germany
sandra.noubours@fkie.fraunhofer.de

Fast Realization of Automatic Translation Systems for New Mission-Relevant Languages

Dr. Matthias Hecking

Sandra Noubours

Fraunhofer Institute for Communication,
Information Processing and Ergonomics FKIE
Neuenahrer Straße 20, 53343 Wachtberg, Germany
matthias.hecking@fkie.fraunhofer.de
sandra.noubours@fkie.fraunhofer.de

Abstract

Documents written in foreign languages are important in preparation of military operations or during these operations. Often not enough human translators and also no automatic tools for translating are available. This problem increases if the documents are written in less-learned languages. Then even less translators and no automatic translation tools are at hand. Therefore, there are little possibilities for the military analyst to get the information out of these foreign documents. If new deployments are planned or new languages for intelligence purposes are identified, the question arises, how the military can react agile to this language problem. In this paper we propose a concept to improve this situation. The main points of the concept include reducing the expectations for the quality of the translation and using the approach of statistical machine translation to rapidly produce new translation systems. The presented concept also includes organizational aspects, but the main focus in this paper is on technical aspects. As an example we show a translation system for the language pair Dari – German. We report about the corpus construction and the experiments to improve the translation system.

1. Introduction

In preparation for military operations or during these operations *documents written in foreign languages* might occur. And the information in the documents might be of great value for the military analyst. There are various ways for automatic information extraction [Hecking, 2011a; Noubours, 2011]. The foreign languages can be an obstacle. It is not a problem to access the content of foreign documents, if enough human translators or appropriate tools for the automatic translations are available. It is a problem if the documents are written in *less-learned languages*. Less-learned means only a few or no human translators are available for translating the documents into the national language of the deploying country (e.g., translating Dari documents into German). But less-learned often also means that there is no economic interest in the industry for building automatic translation systems. Therefore, there are little possibilities for the military analyst to get the information out of these foreign documents.

If new deployments are planned or new languages for intelligence purposes are identified, the question arises, how the military system can react agile to this language problem. In this paper we propose a concept to improve this situation. The main points of the concept include reducing the expectations of the quality of the translation and using the approach of *statistical machine translation* (SMT) to rapidly produce new translation systems. Both points are interrelated. It is only possible to rapidly construct SMT systems if for the analyst *rough translations* (gisting) are sufficient. If the volume of documents to be translated is high, the rough translation is sufficient to identify those documents which should be translated by

human translators. As an example, we describe how we used this approach to set up an SMT system for the language pair Dari – German. The presented concept also includes organizational aspects, but the main focus in this paper is on technical aspects.

The paper is structured as follows. In Section 2 we describe in more detail the basic problem. Section 3 contains the description of the concept for the rapid realization of translation systems. The technical aspect of the framework is presented in more detail in Section 4, where the SMT system for the language pair Dari – German is presented. The paper ends with the conclusion in Section 5.

2. Rough translation

The new missions of the German Federal Armed Forces (Bundeswehr) cause the necessity to analyze large quantities of *documents written in foreign languages*. During the preparation of a new mission or during the mission the content of this written material (web pages, documents ...) might be of great value for the military analyst. According to the “Ethnologue: Languages of the World“ [Lewis, 2012] there exist 6,909 living languages and a lot of dialects. The languages are very diverse in their writing system, lexis, phonetics, and grammar. Figure 1 gives a survey of the different language families of the world¹.

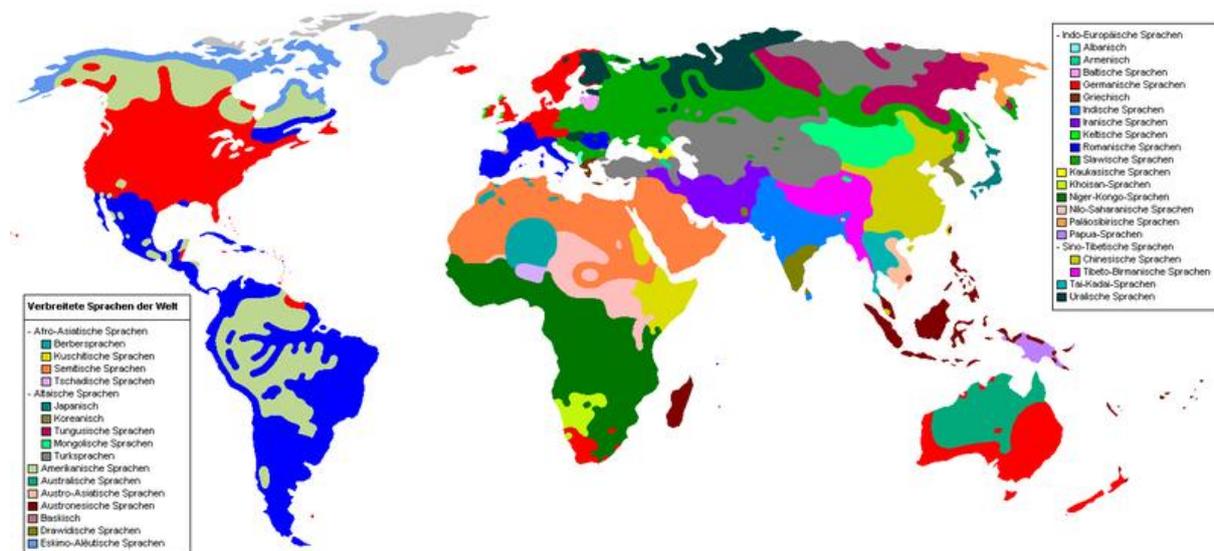


Figure 1. Language families and languages of the world

There are various information extraction techniques to access the content of documents written in foreign languages [Hecking, 2011a; Noubours, 2011]. One important point is always the translation step from the foreign language into the own language. If the foreign language is a less-learned language this means that only a few or no human translators are available for translating the documents into the national language of the deploying country (e.g., translating Dari documents into German). But less-learned often also indicates that there is no economic interest in the industry for building automatic translation systems. Therefore, there are little possibilities for the military analyst to get the information out of these foreign documents. The situation is worsening if an abundance of documents must be translated.

¹ Source: http://diq.wikipedia.org/wiki/File:Sprachen_der_Welt.png

If new deployments are planned or new languages for intelligence purposes are identified, the question arises, how military organizations can react agile to this language problem. Possible solutions might be:

- 1. Hire enough human translators for the language. For less-learned languages this is often not possible. Also the appropriate security clearance must be taken into account.
- 2. Use commercial available automatic systems to translate all documents. Often there is no economic interest in “small languages” and therefore no automatic system available.
- 3. Try to find an approach in which the shortage of human translators and automatic systems is alleviated.

The concept we propose follows the third solution.

Machine translation

Machine translation (MT) is the complete automatic translation of text (our focus) or speech from one natural (source) language to another (target language) while preserving the meaning. It should not be confused with computer-aided translation used by human translators (translation memories).

There are various technical approaches for MT. *Word-to-word translation* gives only a by-word translation. No grammar (e.g., change of word order) is taken into account. The *transfer-based approach* uses different types of analysis (e.g., morphological and grammatical analysis) on the source language side to produce a more abstract description of the text. This description is then transferred into an abstract description of the target language. And from this the translated text is generated. In the *statistical approach* the translation system is generated from bilingual texts (parallel corpus), i.e., where for each given foreign sentence the corresponding sentence of the target language is given (more on this in Section 4).

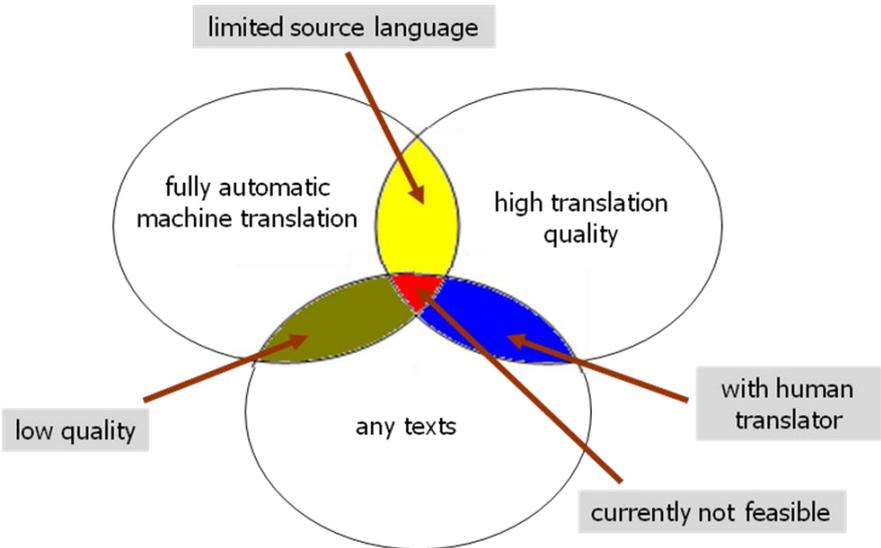


Figure 2. Different types of translation

Several commercial MT systems exist for widely used languages. Online services are also accessible (e.g., Google translator). For less-learned languages they deliver bad quality. And classified texts can't be translated by these services.

A *rough translation* of a foreign text (gisting) doesn't deliver a high quality translation. The translation might contain wrong translated words, grammar errors etc. Depending on the use of the translated texts the rough translation might be sufficient, e.g. if from a lot of documents the relevant ones should be selected for internal information purposes.

In Figure 2² different types of using machine translation is depicted. There are three parameters: the *fully automatic machine translation*, the *high translation quality*, and *any text* to translate. The intersections show different possibilities. If we want to have fully automatic MT on any text this is only achievable if we are willing to accept a low quality (rough translation). If our goal is to get high translation quality fully automatic we can reach this only by restricting the source language (limited vocabulary and phrase structures) of the texts. High quality for arbitrary texts is achievable only by human translators. The intersection of all three parameters is not feasible with the current MT technology.

For new mission-relevant languages any text written in this language might be of interest. Therefore, we have to accept the low quality of the translation. According to Figure 2 this is doable by fully automatic machine translation systems. But, if we want to adapt agile to new "language-situations" this is only possible if the systems for fully automatic translations can be constructed rapidly.

Examples for missions in which unusual language occur are, e.g., the ISAF mission (languages: Farsi/Dari, Pashto) or the EUFOR RD Congo mission from 2006. In the Democratic Republic of Congo 242 languages are used. The official language is French. National languages are Kikongo (Kituba), Lingala, Tshiluba and Swahili. The Congo mission lasted only for four month.

Examples for rough translation systems

A very interesting example of creating a translation system very fast is the *Haitian Creole to/from English* system [Lewis, 2010]. After the disaster in Haiti researchers from Microsoft Research realized a first version of this SMT translation system for texts within four and a half day. The idea was to have a system for translating emergency relief documents, medical documents, SMS text messages, and common phrases and expressions. Haitian Creole is a resource poor language. So, one task during the development was the identification of available data sources. The source data was cleaned-up by bilingual speakers. For this, experts in Haitian Creole joined the team. Before using the SMT software various preprocessing had to be done. One reason for this was, e.g., that Creole is fairly "young" as a written language and in its early stage of orthographic standardization. After the first version of the system and months later the system had around 150,000 segments (phrases) for training data. For both language directions an SMT translation system is now available. Lewis concludes that MT might be a crucial component in crisis situations and can be developed rapidly. He also concludes that SMT systems can be developed by putting together data from a variety of sources and by engaging native speakers and a broader community (crowdsourcing) to assist in the effort.

² The graphic is from the talk "Machine Translation II" given by Harold Somers, School of CS, University of Manchester.

A military operational example for a rough translation system is the *Forward Area Language Converter* (FALCon; cf. [Holland, 1999], [Voss, 2000], [Olive, 2011]). The FALCon is a notebook-based translation system. It can scan documents and does OCR (optical character recognition) from the resulting images to text. The text is then translated into English and the English text can be searched for keywords. For the translations an off the shelf product was used. The intention was to give document collectors in the field the possibility to quickly scan the documents for mission relevance. Only the relevant documents are then given to human translators. The FALCon system was developed by the U.S. Army Research Laboratory (ARL) starting around 1995. The system was used during the Haiti (1995) and the Bosnia (1997) mission. Feedback from field use in Bosnia [Holland, 1999] showed that the system was indeed sufficient for document screening and that it reduced the workload of linguists.

3. Concept for the rapid realization of translation systems

We present here a concept to overcome the described problem. The concept contains the following elements to react agile to the need of new translation systems:

1. Reduce the expectations concerning the *quality* of the automatic translation to rough translation,
2. use the approach of *statistical machine translation* to come up very fast with a new translation system,
3. build up a *team of scientists* who are up-to-date to the SMT technology and to the corresponding scientific field and who are responsible to create very fast new versions of the translation system,
4. create an military operational centralized automatic *translation service* for the military users via any military intranet, and
5. make sure that the operational staff tightly works together with the team of scientists.

As we already explained, high quality translation for arbitrary texts is not reachable with the current translation technology. Therefore, the expectations concerning the quality of translations have to be reduced (1st element). Only rough translations (gisting) are feasible for any text. The SMT approach is the technical approach through which - at the moment - translation systems can be generated very fast (2nd element). The SMT systems are developed in a generate-and-evaluate cycle. During each cycle the scientists use the newest training material (e.g., collected during the use of the operational system, produced by crowd-sourcing or corpus linguists) and the newest research results to generate a new version of the translation system (3rd element). To make sure that each user of a specific language pair has access to the most up-to-date translation system there should be a centralized automatic translation service accessible via a military intranet (4th element). The operational people of the translation service should work tightly coupled with the team of scientists to make sure that the newest training material is used during the development of the new version of the SMT system. Also, the needs of the actual users can be communicated faster to the scientists. The scientists are also responsible for applying the latest research results during the developments (5th element).

Crucial is the close cooperation of the operating people and the developer of the MT system. This reduces the turnaround time for constructing the newest version of the MT system. It also assures that the needs of the military end users are taken into account. Beside specialists for SMT also corpus linguists are important. They should be members of the scientists group.

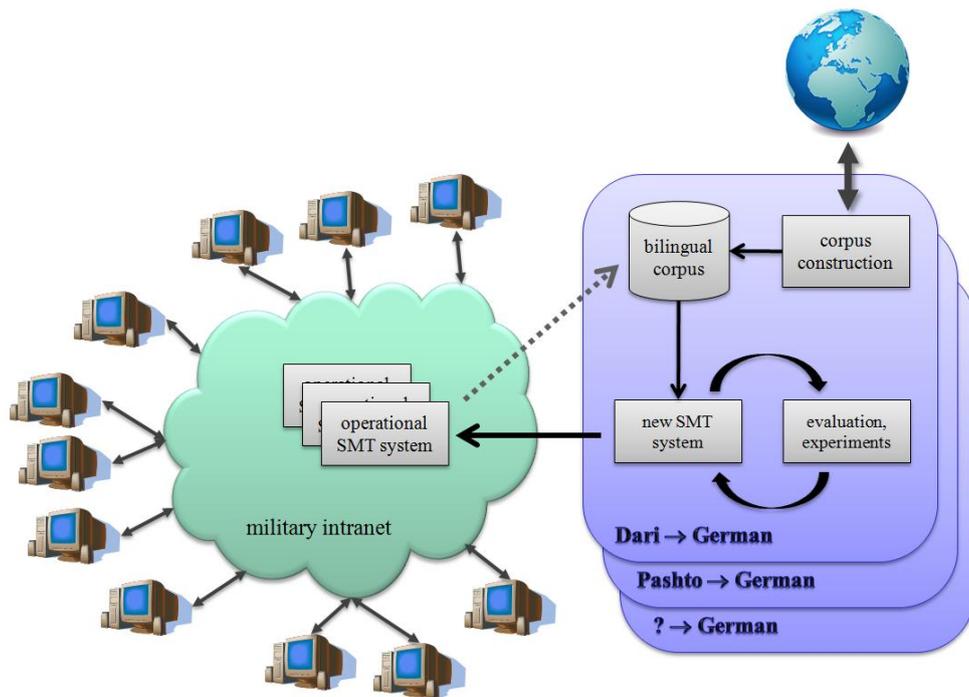


Figure 3. Concept for the rapid realization of translation systems

In Figure 3 the concept is sketched. For each language pair an SMT system is constructed by experiments and evaluations. To get an appropriated bilingual corpus all possible sources (“the world”) must be used. If a new SMT system is ready it is pushed into the operational environment (military intranet). Feedback is given from the operational SMT systems mainly to the corpus builder.

4. Language pair Dari – German as an example

The approach of SMT (2nd element in our concept) is crucial for the success of the concept. Therefore, we build up a small infrastructure (small team of scientists, computer cluster, procedures, software ...) to show that we can come up fast with a new SMT system for a language pair relevant for the Bundeswehr.

ISAF-MT project

One of the objectives of the ISAF-MT project is to show that the SMT approach is appropriate for building translation systems for military-relevant “unusual” language pairs, in our case for *Dari – German*. Dari is spoken and written in the mission area of the Bundeswehr in Afghanistan. Our previous work on translation focused on simple word-to-word translations for Dari [Hecking, 2008] and Tajik [Hecking, 2010a] to German. We also made some experiences with SMT for the language pair Dari – English [de Bond, 2009].

Our research project *Machine Translation for ISAF Forces (ISAF-MT)* is a German – U.S. cooperation project. On the U.S. side the Air Force Research Laboratory (AFRL, Wright-Patterson Air Force Base) is responsible. In the ISAF-MT project our objectives are to build up bilingual corpora and linguistic tools and to construct through SMT technology Dari – German translation systems.

Our project also proved that after building up the SMT infrastructure (computer cluster, procedures for building corpora, etc.) a translation system for rough translation can be produced rapidly.

Dari³

In Afghanistan, due to the diversity of ethnic groups, a lot of different languages are spoken. The mother tongue of about half of the approximately 22 million Afghans is *Dari*. It is beside Pashto the second official language of Afghanistan. Dari is spoken in the center and in the northern part of Afghanistan. Various dialects exist. The default is the pronunciation of Kabul. Dari is an Indo-European language. The writing system was adopted from Arabic. Four additional characters have been added to the 28 Arabic characters to handle sounds not existing in Arabic. Dari is a right-to-left language, does not distinguish between uppercase and lowercase, and does not write short vowels. The language is characterized by a syntactic word order SOV (subject-object -verb) and has an affixal morphology. Dari is very closely related to the modern Iranian Persian, called *Farsi*. Another closely related language is *Tajik* spoken in Tajikistan the northern neighboring state of Afghanistan. However, the Tajik uses Cyrillic characters.

Introduction into SMT

The goal of statistical machine translation is to find the “most likely” translation for a given source language sentence [Koehn, 2010]. For the source language sentence f the target language sentence e is selected that maximizes the probability $p(e|f)$. With the Bayes Rule we get:

$$\arg \max_e p(e|f) = \arg \max_e \frac{p(e)p(f|e)}{p(f)}$$

This model defines the best translation through two components: the *translation model* $p(f|e)$ and a *language model* $p(e)$. Intuitively the translation model (example excerpt in Figure 4) retains the content of the source language sentence and the language model (example excerpt in Figure 5) makes sure that the target language sentence is well-formed. The probabilities are determined (trained) by machine learning algorithms using large text corpora. For the translation model large *bilingual parallel corpora* are needed. These are collections of translated texts in both the source and the target language. During the training of the language model only a *monolingual corpus* of the target language is required.

At first, the training of an SMT system needs the preprocessing of the corpus. This is a technical problem but might be time-consuming. After this, the translation and language model is trained. The translation model consists of various submodels (e.g., for alignment and reordering) with its own probability values. These are the numbers in Figure 4. Therefore, in the tuning step the parameters of these probabilities (and that of the language model) are adjusted to get the best translation of the whole system.

The component which is used to translate texts – the ready to use SMT system - is called *decoder*. It contains a search algorithm to determine for a given source language sentence the best translation according to both models. In general, several translations are produced with their probabilities.

³ http://en.wikipedia.org/wiki/Dari_%28Persian%29

The SMT system can also be build to give translations of whole phrases und to take into account linguistic information on both language sides (e.g., morphological or part-of-speech information).

154229	، ساکت باتنید ، , sei still !	0.197548	0.00202913	0.197548	0.00328463		1 1
154230	، ساکت باتنید ، , sei still	0.197548	0.00331125	0.197548	0.00420655		1 1
154231	، سایت Welt . online نویسد : درمطلبی با عنوان می نویسد	Quelle	Welt	online	am 20. Oktober 2008	:	0.0282211
154232	، سایت Welt . online نویسد : درمطلبی با عنوان می نویسد	Quelle	Welt	online	am 20. Oktober 2008		0.0282211 4.
154233	، سایت های Stationierung von	0.00581022	3.81773e-06	0.0246934	7.41779e-06		34 8
154234	، سایت های Stationierung	0.00548743	3.81773e-06	0.0246934	0.0006259		36 8
154235	، سایت های begleitet von der Stationierung von	0.00731657	3.81773e-06	0.0246934	2.60816e-13		27 8
154236	، سایت های begleitet von der Stationierung	0.00731657	3.81773e-06	0.0246934	2.20072e-11		27 8
154237	، سایت های der Stationierung von	0.00731657	3.81773e-06	0.0246934	4.73274e-07		27 8
154238	، سایت های der Stationierung	0.00731657	3.81773e-06	0.0246934	3.9934e-05		27 8

Figure 4. Excerpt from the ISAF-MT translation model

-0.9569345	für die Dauer des	-0.140566
-0.3309343	für die Dauer von	0.0009403285
-1.329289	nur die Dauer von	-0.006186663
-0.4907208	oder die Dauer der	-0.006186663
-0.4907208	und die Dauer der	-0.006186665
-0.4907208	war die Dauer der	-0.2274503

Figure 5. Excerpt from the ISAF-MT language model

The evaluation of SMT systems is based on a manually checked bilingual corpus (around 1,000 phrases). After translating the sentences the results are checked automatically against the given reference translations. This comparison is the basis for computing various evaluation metrics. These metrics try to measure without a human in the loop how “good” the translation is. Various metrics are use, e.g., BLEU, METEOR or TER. See [Koehn, 2010] for more information on these metrics.

The most widely used toolbox for constructing SMT systems is *Moses* [Koehn, 2007]. Moses is an open-source project and has a very large user and developer community. This ensures that the latest concepts and techniques are available for developers of SMT systems. The ISAF-MT project also uses Moses.

The quality of the SMT system is determined by the amount of bilingual training material, whether the topic of the foreign texts varies a lot, the coverage of the training data etc.

The Dari – German corpus

SMT systems are generated from bilingual corpora, i.e., a lot of equal texts in the source and the target language. The texts are sentence-aligned. In Figure 6 an example from our *Dari – German corpus* is shown. In each line of both texts the corresponding sentence is stored.

1	دوی نیروهای امنیتی افغان ثابت کردند که توانایی تامین فرقه جدید ۱۱۱ امنیت کابل را تقویت می بخشد.	1	Die 111. Division unterstützt die Sicherheitskräfte in Kabul. Nachdem afgha
2	وزارت دفاع ملی در اخیر ماه می فرقه جدید ۱۱۱ اردوی ملی افغان را در کابل تاسیس نمود.	2	Ende Mai wurde die 111. Division vom afghanischen Verteidigungsministe
3	این فرقه با همکاری نزدیک با دیگر نیروهای ملی امنیت افغان چون یولیس و امنیت کار خواهد نمود.	3	Die Division wird in Kooperation mit anderen afghanischen Sicherheitskrä
4	پتخت می باشد، فرقه ۱۱۱ مسئولیت امنیت شهر را زمانی بدوش خواهد گرفت که کابل را خطر بزرگی تهدید نماید	4	Während die afghanische Nationalgarde hauptverantwortlich für die Vertei
5	دم خدمت خواهیم نمود	5	"Ich verspreche der afghanischen Regierung, dass wir unserem Volk unte
6	ء و پنج کتک می باشد	6	Kabuls 111. Division, die ihr Training vor sechs Monate begonnen hat, bes
7	کنندگان مجهز می شوند	7	Die Soldaten besuchen ein zweimonatiges Training und ihre modernen Wi
8	نیز قواء مسلح وزارت	8	Die Gründung der neuen Division war ein großer Schritt in der weiteren Ei
9	زی را فراهم نموده اند	9	"Die afghanische Armee hat mehr Sicherheit in die afghanischen Provinze
10	ی خارجی کار می کنند	10	Weiterhin kündigte er eine neue Militärinitiative an, die aus Polizeibeamt
11	مه بین المللی شده است	11	Die letzten Vorfälle, in denen Zivilisten aus Versehen ums Leben gekomme
12	قداست می باشد بهبود	12	Die sieben Jahre alte afghanische Armee besteht mittlerweile aus 90.000

Figure 6. Example from the Dari – German corpus

To generate an SMT system the bilingual text corpus should be as large as possible. For “big” languages and common language pairs this is often not a problem. For example, for English and French there is a lot of material available or can be produced from the internet. This is not the case for a less-resource language like Dari and it’s obviously not the case for the unusual language pair Dari – German. This means, that our focus lays on the production of the corpus by extracting parallel texts from the internet and from various military sources and also by producing own bilingual documents by native Dari speakers. For the last task we looked for entries with the topics “military” and “terrorism” in the online versions from Sada-e-Azadi⁴, Pajhwok⁵ und Kokchapress⁶. Various other sources were also used. At the moment the corpus has around 27,000 sentences. We also integrated a military dictionary Dari to German. It contains 71,798 entries.

The experiments

The overall objective of the experiments is to find improvements of the translation model (and its submodels) and correct weights of the parameter in the models that maximizes the probability of produced translated sentences. The following parameters were considered: *size and quality of the tuning set, normalization, compound splitting, reversal of sign and word order, alignment heuristics, maximum phrase length, phrase scoring, reordering, inclusion of part-of-speech (POS), inclusion of lemma, and inclusion of word stems*. A comprehensive description of the experiments can be found in [Hecking, 2010b] and [Hecking, 2011b].

The different experiments run on a small computer cluster. The experiments are controlled and distributed on the network by the Oracle Grid Engine/Sun Grid Engine (SGE). SMT performance is measured in BLEU, METEOR and TER. For better models the BLEU and METEOR scores must increase and the TER value must decrease. In the following, some of the experiments are explained in more detail. The baseline system comprises certain configurations of the translation and language model which are used as a starting point for the change of the considered parameter. Please note, that the baseline systems in the different

⁴ <http://sada-e-azadi.net> (24.1.2012)

⁵ <http://www.pajhwok.com> (24.1.2012)

⁶ <http://kokchapress.com> (24.1.2012)

experiments are not the same. The experiments describe different development levels of the Dari – German translation system.

The German language has a lot of *compound words*. These are words consisting of more than one stem, e.g., “darkroom”. Often, these result in problems during automatic language analysis. In the experiments *Compound splitting* the compounds were split by the *Moses Compound Splitter* into single words. This was done for both languages. The results are:

Compound splitting	BLEU	METEOR	TER
baseline	7,55	15,60	8,93
input (dari)	7,14	15,54	9,02
output (deutsch)	6,65	14,98	10,50
input (dari) & output (deutsch)	7,14	15,60	9,02

Compound splitting results in a decline in the scores. Contrary to the expectation compound splitting does not increase the quality of the translation.

Dari has a *right-to-left writing system* and the verb is located at the end of the sentence. This is a notable difference to German. In the experiments *Reversal of sign and word order* each line of the Dari side was reversed sign by sign and word by word. The results are:

Reversal	BLEU	METEOR	TER
baseline	11,10	16,55	8,87
reverse words	10,03	15,48	9,41
reverse string	10,07	15,50	9,38

The results show that reversing sign and word order leads to a deterioration of the evaluation scores. Simple reversing is not an appropriate way of handling the direction in different writing systems.

Lemmas are chosen word forms for lexicon entries. They are abstractions of the various possible word forms, e.g. verb conjugations. In the experiment *Inclusion of lemma* it was supposed to show whether lemma annotation increases BLEU score or not. Lemma information for both languages was included. To integrate this information so-called *factored translation models* [Koehn, 2009] were used. The German side of the corpus was annotated by the Mate-Tools package [Björkelund, 2011] and the Dari side was processed by an internal tool. In the following table the results for the experiments (excerpt) with various possibilities to use the lemma information are given:

Inclusion of lemma	BLEU	METEOR	TER
Baseline	5,34	12,14	9,11
align: lemma => lemma	5,77	12,60	9,03
align: word => lemma	5,60	12,14	9,02
align: word + lemma => lemma	5,61	12,23	9,02
trans: lemma => word + lemma	5,37	12,18	9,06
trans: word => word + lemma	5,45	12,18	9,09
trans: word + lemma => word	5,35	12,15	9,08
trans: word + lemma => word + lemma	5,36	12,13	9,09

reord: lemma => word	5,40	12,32	9,07
reord: word + lemma => word	5,42	12,26	9,09

We found that alignment, training and reordering (different substeps of the translation model) can benefit from the inclusion of lemma. The biggest improvement is achieved by a word alignment “lemma => lemma”.

In Figure 7 real examples of translated phrases together with the expected translation (reference) are given. They show different degrees of semantic equality.

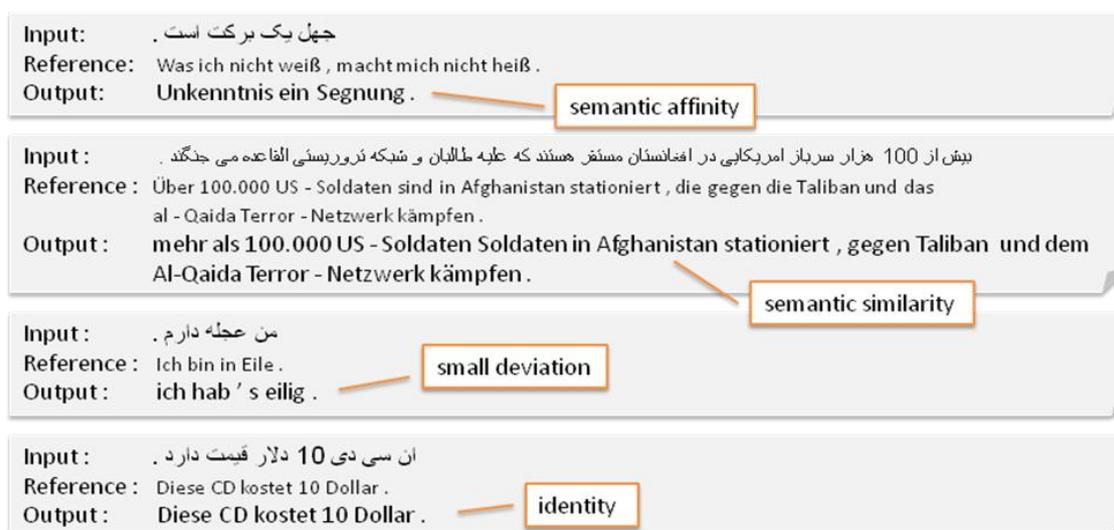


Figure 7. Example outputs of the ISAF-MT translation system

We were successful in realizing an SMT system of a language pair relevant for the needs of our Armed Forces. We have not done this within 4 and a half day, like in the Haitian case, due to corpora availability but we were able to do this in a couple of month. The ISAF-MT translation system is still under further development and also under operational test.

5. Conclusion

Understanding documents written in foreign languages are important in preparation of military operations or during these operations. If there are not enough translators and/or it is a less-learned language there are little possibilities for the military analyst to get the information out of these foreign documents. To overcome this unsatisfying situation we propose a concept in this paper. The main points of the concept include reducing the expectations for the quality of the translation and using the approach of statistical machine translation to rapidly produce new translation systems. As an example of using the SMT technology we show a translation system for the language pair Dari – German, a language pair which is important for the German Federal Armed Forces. We report about the corpus construction and the experiments to improve the translation system.

The SMT technology is under heavy development. It is supposed that this development will lead to better translation quality. Research also starts handling the problem of less-resource languages and considering semantic information in the translation model. In the ISAF-MT project we will incorporate these new research results in our system.

References

- [Björkelund, 2011] A. Björkelund, B. Bohnet. *Mate-tools - Tools for Natural Language Analysis, Generation and Machine Learning*. <http://code.google.com/p/mate-tools/> (12.12.2011).
- [de Bond, 2009] C. de Bond. *Statistische maschinelle Übersetzung von Dari nach Englisch unter Einbeziehung linguistischer Faktoren bei begrenztem Trainingsmaterial*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), Wachtberg, Germany, FKIE-Bericht Nr. 175, 2009.
- [Hecking, 2008] M. Hecking, C. Schwerdt. *Multilingual Information Extraction for Intelligence Purposes*. In: Proceedings of the 13th International Command and Control Research and Technology Symposium (ICCRTS), Bellevue, WA, U.S.A., 2008.
- [Hecking, 2010a] M. Hecking, T. Sarmina-Baneviciene. *A Tajik Extension of the Multilingual Information Extraction System ZENON*. Proceedings of the 15th International Command and Control Research and Technology Symposium (ICCRTS), Santa Monica, CA, U.S.A., June 2010.
- [Hecking, 2010b] M. Hecking, S. Noubours. *Machine Translation for ISAF Forces (ISAF-MT)*. Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie FKIE, Wachtberg, Germany, Dokumentation zum Forschungsvorhaben E/IB1S/AA166/9F008, 31. Dezember 2010 (in German).
- [Hecking, 2011a] M. Hecking, A. Wotzlaw, R. Coote. *Multilingual Content Extraction Extended with Background Knowledge for Military Intelligence*. In: Proceedings of the 16th International Command and Control Research and Technology Symposium (ICCRTS), Québec City, Québec, Canada, 2011.
- [Hecking, 2011b] M. Hecking, S. Noubours. *Machine Translation for ISAF Forces (ISAF-MT)*. Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie FKIE, Wachtberg, Germany, Dokumentation zum Forschungsvorhaben E/IB1S/AA166/9F008, 31. Dezember 2011 (in German).
- [Holland, 1999] V.M. Holland, C.D. Schlesiger. *High-Mobility Machine Translation for a Battlefield Environment*. Paper presented at the RTO SCI Symposium on “The Application of Information Technologies (Computer Science) to Mission Systems”, held in Monterey, California, USA, 1998.
- [Koehn, 2007] 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, E. Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.
- [Koehn, 2009] Koehn, P., Hoang, H. *Factored Translation Models*. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational (EMNLP 2007), Prague, Czech Republic, 2007.
- [Koehn, 2010] Koehn, P. *Statistical Machine Translation*. Cambridge, UK: University Press, 2010.
- [Lewis, 2010] W. D. Lewis. *Haitian Creole: How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes*. 14th Annual Conference of the European Association for Machine Translation (EAMT), St Raphael, France, 2010.
- [Lewis, 2012] M. Paul Lewis (ed.). *Ethnologue: Languages of the World*. Sixteenth edition. Dallas, Tex.: SIL International, 2009. Online version: <http://www.ethnologue.com/>, (19.1.2012).

- [Noubours, 2011] S. Noubours, M. Hecking. *Semantic Analysis of Military Relevant Texts for Intelligence Purposes*. In: Proceedings of the 16th International Command and Control Research and Technology Symposium (ICCRTS), Québec City, Québec, Canada, 2011.
- [Olive, 2011] J. Olive, C. Christianson, J. McCary (eds.). *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer New York, 2011.
- [Voss, 2000] C. R. Voss, C. Van Ess-Dykema. *When is an embedded MT system "good enough" for filtering?* In: Proceedings of ANLP-NAACL 2000 Workshop: Embedded Machine Translation Systems (EmbedMT '00), Seattle, WA, U.S.A., 2000.