

Learning, Forgetting and Remembering: Statistical Support for Rule-Based MT

Oliver Streiter, Leonid L. Iomdin, Munpyo Hong and Ute Hauck
IAI, Institute for Applied Information Sciences
ITTP, Institute for Information Transmission Problems
Martin-Luther Str. 14, D-66111 Saarbrücken, Germany
Bol'shoi Karetnyj Pereulok 19, Moscow, 101447, Russia
oliver,munpyo,ute@iai.uni-sb.de, iomdin,oliver@proling.iitp.ru

Abstract

The paper describes the incorporation of statistical knowledge into two different Rule-Based MT (RBMT) systems. In earlier experiments, these systems were linked with Memory-Base MT components, so that by now the translation process is supported by three MT paradigms. The paper concentrates on the acquisition on rich, informative, balanced, and up-to-date statistical data from monolingual and parallel corpora and on ways of using these data in RBMT systems. The authors keep their pledge of a systematic investigation of the linkage of different MT paradigms aimed at improving the quality of translation.

1 Introduction

For a long period, the rule-based approach has been the only strategy pursued by researchers in the field of MT. The appearance of Translation Memories (TMs), Statistics-Based MT and Example-Based MT (EBMT) has changed the situation by shifting the acquisition of data away from the linguistically inspired human rule writer to the machine itself, which acquires translation knowledge (a) faster than its human counterpart, (b) in greater quantities, (c) more reliable with respect to statistical information, (d) related to larger text chunks than intuitively felt necessary by the human counterpart and (e) in a format which may be used directly by the machine. After a period of experiments and discussions it has become clear, that none of the approaches in their isolated form will solve the problem of MT within a reasonable time (Som98). It is equally unlikely that a new, "ideal" approach may be proposed and implemented on a sizeable scale in the foreseeable future. However, it could be shown, that progress can be achieved by combining the strengths of different approaches (FN94) (N97).

Although the combination of the graphematic output of different MT engines as pursued in (FN94) is not trivial, we are convinced that the different engines have to interact on intermediate, thus richer, representations. Otherwise, the strong sides of one engine cannot be fully integrated and compensate for the weak sides of a second engine. Long-distance dependencies or variations in word order cannot be repaired through the combination of different incomplete translation. If however, one engine can handle word order and the second engine the translation of idiomatic expression and both can exchange their knowledge, considerable improvement can be achieved.

In previous research we showed how RBMT and EBMT can interact (CIS98), (CPS99). In this paper we investigate how simple statistical data can be used within the framework of an RBMT system. Besides the attempts to integrate statistics into general rule-based NLP-frameworks, e.g. (Res92), it has been shown, that a statistically enriched RBMT systems can handle collocational phenomena; thus to find the most likely translation of *match* in the context of *fire* or *ball*. These approaches are based on monolingual corpora (Nom91) or one independent corpora of source and target language (DM92).

In the two experiments described here, we work on corpora classified with respect to a subject domain. The aim of the experiments is to extract simple statistical information out of the corpora and by this measure to tune the system automatically to a specific subject domain. In the first experiment, we used **monolingual** corpora to extract statistical information. The obtained numerical data are transformed into a rank-order of preferences and compiled into the lexicon of the RBMT system. In the second experiment, we used **parallel** corpora in order to rate the translation hypotheses created by the RBMT system with respect to their likelihood of occurrence in a text belonging to a given subject domain. The rated hypotheses are transformed into a rule format and compiled into the MT lexicon. Both experiments have been realized on two different RBMT system.

2 Experiment 1: Monolingual Corpora Support

Since large parallel corpora may be hard to find for some language pairs and subject domains, we first investigated the possible use of statistical data drawn from monolingual corpora for RBMT.

2.1 Collection of Classified Corpora

In order to obtain up-to-date word frequencies we collect corpora via a web-based translation service. MT users submit texts for translation and classify them with respect to a set of subject fields (e.g. medicine, biology, society, sports, economy, computing etc). This 'automatic' acquisition and classification ensures a continuous growth and actualisation of the data. These texts are automatically submitted to a morphological analyser and the output of the latter is subjected to statistical analysis and to the translation system. The statistical data obtained can be used in one of the subsequent translation sessions, as soon as they have been compiled off-line into the lexicon of the RBMT system.

2.2 Morphology-Mediated Word Frequencies

The morphological analysis of the texts yields a description of words with respect to their inflection, derivation, and compounding. All these pieces of information are not only necessary for the task of translation but also helpful in the acquisition of statistical data. If statistical processing builds upon morphological representations instead of the occurrence of words in a text, we obtain data which are informative (having a high predictive power), rich (after the morphological processing we have more data than before) and balanced (we incorporate data on discontinuous words).

2.2.1 Inflection

It goes without saying that the reduction of as few as two inflected forms to one underlying lemma renders the data more informative. For example, if the source text contains two word forms, e.g. *players* and *player*, the resulting statistics will count two times *player*. It is even more important to use information on inflection if the source language has rich morphology.

2.2.2 Compounding

If morphological analysis of compounding is used to calculate word frequencies, we obtain information not only about the compound word itself, but also about its components. Thus, the German word *Fussballspieler* 'soccer player' supplies the statistics with frequencies for *Fussballspieler*, *Fussball* 'soccer' and *Spieler* 'player'.

2.2.3 Derivation

The analysis of derivation may be used in a similar way. It may help, for example, to acquire word frequencies of words that do not appear as one graphematic unit in the text but are separated by other words. Without a complicated syntactic analysis, such variants are necessarily underrepresented, e.g. in the German sentence *Der Spieler spielt den Ball ab* (*The player is feeding the ball*) the frequency which should be ascribed to *abspielen* 'feed', 'pass forward' is ascribed to *spielen* 'play'. At the moment, we cannot change this unhappy situation but try to compensate for it with the help of derivational morphology. Thus, if the German word *Abspiel* 'feeding', an irregular verb-to-noun derivation, appears in the text, it can be counted not only as *Abspiel* but also as *abspielen*. Thus, even though the un-prefixed verbs may be overrepresented, we obtain significant data about the prefixed verbs.

2.2.4 Examples

A Perl program identifies within the morphological analysis the relevant substrings (the base of the derivation, the parts of the compounds) and performs the counting. Some simple examples in (1) illustrate how several frequencies are obtained out of one word.

analysed word	counted word forms
Fussballspieler (soccer player)	spielen, fussballspieler, spieler, fussball
Abspiel (feeding)	abspiel, abspielen
Realisierbarkeit (implementability)	realisieren, realisierbar, realisierbarkeit

Figure 1: Examples of morphology-mediated word frequencies

2.3 Compiling Frequencies into the Lexicon

In the CAT2 RBMT system used in the first experiment, a lexical entry contains, beside its monolingual descriptors, links to lexical entries of other languages. The German lexical entry *allgemein* is linked, among others, to a set of English lexical entries, one of them shown in figure 2.

During one of the compilation steps this representation is modified in such a way that the word frequencies of different subject fields (e.g. sports, medicine, economics, computing, etc.) are added in the form of preferences with the leftmost translation

```
{l=allgemein,...,en={t=(common;general;generic;overall;public;universal)}}
```

```
{l=general,...,de={t=(allgemein;general;generell;gesamt;global)}}
```

Figure 2: Simplified entries of *allgemein* and *general*

being the most frequent and thus most preferable for a subject domain. As the examples show, *general* contains references to more subject domains than the German entry *allgemein*. This is accounted for by the fact that such references are only supplied if they are attested for the given subject domain. Since, for instance, none of the English words cited in *allgemein* were found in medical texts, no references to the domain of medicine could be made.

```
{l=allgemein,...
  en=({t=(overall;universal;generic;general),dmn=sports}
      ;{t=(general;overall;generic;universal),dmn=society}
      ;{t=(general;universal;overall;generic),dmn=common}
      ;{t=(general;overall;generic;universal),dmn=comp_sc})}

{l=general,...
  de=({t=(gesamt;global;generell;general;allgemein),dmn=sports}
      ;{t=(gesamt;allgemein;generell;global;general),dmn=society}
      ;{t=(gesamt;allgemein;generell;global;general),dmn=common}
      ;{t=(allgemein;global;gesamt;generell;general),dmn=medicine}
      ;{t=(allgemein;gesamt;general;global;generell),dmn=economy}
      ;{t=(gesamt;allgemein;generell;general;global),dmn=comp_sc})}
```

Figure 3: Enriched entries of *allgemein* and *general*

2.4 Source Language Analysis

For a language with many homographs and poor morphology, simple word frequencies may be of limited use for the distinction between homographs. The English word *match*, for example, has, in addition to the homograph verb, at least three nominal readings (cf. figure 4): the verb-to-noun derivation of the verb *to match* (a), the thing used to make fire (b), and the event in which players participate (c). If the word frequencies of the respective (German) translations are taken into consideration, we eventually can distinguish them by reference to the subject fields. For the translation of sport texts, we can thus identify (6) as the most likely candidate for the English analysis of *match*.

In a similar way, if we encounter in a German text the verb form *kommen* 'come', this word will trigger the lexical entries of about 20 German verbs such as *kommen_ab*, *kommen_an*, *kommen_auf* etc. The suggested approach helps to find the correct verb by looking at the translations and whether they are attested for a given subject field. If they are not attested, they are not considered for syntactic analysis.

In order to evaluate the effect on the parse time, we used 50 sample sentences (on the average, 11 words per sentence). These sentences were parsed with and without frequency information of the target word. In 17 of 50 sentences the frequency information reduced the parse time by about 21%. In the remaining 33 sentences no difference was detected. There was no sentence which displayed a worse system performance

- (a) `{l=match,..., de=({t=(uebereinstimmen;matchen),dmn=society}
;{t=(uebereinstimmen;matchen),dmn=common}
;{t=(uebereinstimmen;matchen),dmn=economy}}}`
- (b) `{l=match,...,de=({t=(streichholz)})}`
- (c) `{l=match,...,de=({t=(spiel;partie),dmn=sports}
;{t=(spiel;partie),dmn=society}
;{t=(spiel;partie),dmn=common}
;{t=(spiel;partie),dmn=economy}
;{t=(spiel;partie),dmn=comp_sc})}`

Figure 4: Three nominal concepts of *match*

when supplied with the frequency information. If we consider the whole parse time for all sentences, the statistically endowed system has shown a 13% improvement of the parse time. With an increase of the statistical data two opposite tendencies may be expected. Some homographs may be distinguished and parsing becomes faster, other homographs which are now distinguished may also become more similar if their translations are found in more text types.

	num. of sent.	p-time with freq.	p-time without freq.	improve.
total	50	440 sec	503 sec	12.35%
faster	17	241 sec	304 sec	20.73%
no change	33	199 sec	199 sec	0.00%

Figure 5: Parse time with and without frequency information

2.5 Translation

During translation, the statistics about the target language is used to select the most likely of the possible translations. Thus, the first choice to translate *general* in a medical context is *allgemein* while for sports it is *gesamt*.

Of course, noise produced by homographs of the target language cannot be avoided altogether. For example, the statistics will yield for the German noun *Auto* the preferred translation *car* for all subject fields, except for sports, where it is *coach*. The distortion is accounted for by the second meaning of *coach*, i.e. 'trainer'. In order to reduce such influence we countercheck the target of the translation (here, *coach*) on whether its translations (into German) are attested for the given subject field. Thus, while translating *Auto* into *coach*, we first check whether *coach* has a translation which is attested for the subject field of sports and if it is not, which is the case for the German *Bus*, the system chooses the second translation, which is *car*. If no translation with an appropriate attestation can be found at all the translation is repeated without checking the attestation in a subject field. This fallback position represents the operation of the transfer phase before the integration of the statistical data.

To test our hypothesis, we made a small experiment with 40 German sample sentences, taken from various subject domains, such as sports, economy, medicine, etc.¹

¹We express our gratitude to Rita Nübel for her helps and comments in the evaluation of the

Then we submitted the sentences to the MT-System, first using the frequency information and later without this information. The English translations of the German source texts were evaluated by human translators with the simple criterion whether the quality of the translation is improved or not. The experiment has shown that in 7 of 40 cases the improvement of the translations was achieved with the frequency information and in 2 of 40 cases the results became worsen. In the rest 31 cases, no difference was detected.

total num. of sentences	improved	no difference	worsen
40	7	31	2

Figure 6: Translation quality with and without frequency information

As expected, translations were improved by a better selection of target terms as exemplified in the German sentence *Der Fußballspieler gewann drei Title in diesem Jahr*, translated into *The football player **won** three **titles** in this year* in the 'sports'-context when supported by the statistical information, whilst it was translated into *The football player **gained** three **mastheads** in this year* without the frequency information. In the cases of a deteriorated translation, the system accidentally selected better target items than the statistical data suggested. The sentence *Der alte Präsident traf heute in Albanien ein* is translated without frequency information as *The **old** president arrived **today** in Albania* and with frequency information as *The **aged** president arrived **presently** in Albania*. This deterioration is due to the fact that frequencies of target items are counted independent of the source item (e.g. *presently* is more frequent than *today*). In order to relate source and target items statistically, bilingual corpora have been treated in a second experiment.

3 Experiment 2: Bilingual Corpora Support

The RBMT system used in the second experiment is the ETAP-3 MT System, a short description of which can be found in (CI97).

3.1 Rating Translation Equivalents

The lexica of an RBMT system contain, among other things, a huge set of translation equivalents, from which the system must select the most appropriate ones. In the experiment, we collected a very large list of translation equivalents, using several lexica of the RBMT system. The list, which contained over 130,000 bilingual equivalents (for the most part consisting of one word each) plus information on the part of speech and, in case of ambiguity, the number of the homonym, was then checked against parallel corpora, each associated with a specific subject domain.

Each monolingual text of the parallel corpus is morphologically processed. The morphologically processed texts are then transformed into a string of possible lemmas and homographs are contextually resolved.

Source and target texts are aligned via lexical anchor points. Lexical anchor points are semi-compositional translations consisting of more than one word (e.g. English:

translations.

yellow card, German: *gelbe Karte*, French: *carte en jaune* etc.). Via Dynamic Programming (DP) a path through these anchor points is calculated. Sub-paths and their evaluations are stored so that they do not have to be recalculated. The whole processing is recursive, the construction of the optimal path takes place during backtracking. The whole processing however is sub-optimal, as only two anchor points can be ignored in a row. A 'good' path uses many anchor points and shows modest changes in the slope between the anchor points. After a first alignment new semi-compositional translations are extracted and with these new anchor points the alignment is improved. The process continues cyclically until no new anchor points are found. The final alignment consists of the interpolation between the anchor points of the optimal path.

```
function optimal (list_of_anchor points) {
  if list empty { return (0,medium_slope,last x,last y) }
  else if result stored for list_of_anchor points { return the stored result }
  else {
    optlist1=optimal(FIRST,SECOND,THIRD,...) # recursive calls
    optlist2=optimal(FIRST,THIRD,FOURTH,...) # missing 1 element out
    optlist3=optimal(FIRST,FOURTH,FIFTH,...) # missing 2 elements out
    optimal_list= one of optlist1 or optlist2 or optlist3 # evaluation
    return "optimal value,actual slope,(FIRST,optimal list)" } }
```

Figure 7: A DP Algorithm for a (sub)optimal course of anchor points

With the help of the interpolation we can test for each word of the source language, which of its possible translations actually appeared in the target text close to the position specified by the function. As a result we receive a list of attested translations for a given subject domain and their frequencies. In the following example, the data in each line are read as follows: English lemma, English part of speech, English lexeme number (if any), Russian lemma, Russian part of speech, Russian lexeme number, absolute frequency of the pair, day of the last modification (starting from year's beginning).

3.2 Compiling Frequencies into the Lexicon

As in Experiment 1, the outcome of the statistical analysis is compiled off-line into the lexicon of ETAP-3. The lexicon makes use of different translation fields for different subject domains, which can be created, supplied with information, deleted or modified automatically. Accordingly, lexical entries for which an attested translation has been found in the representative corpus, can be supplemented by a translational equivalent appropriate for the relevant subject domain.

3.3 Source Language Analysis and Transfer

Similar to Experiment 1, it has proven possible to reject a source language analysis on the basis of translation equivalent rating. So, if the syntactic analysis of a sentence identifies an ambiguity at a lexical level so that (at least) one of the alternative word hypotheses bears a translation field of the given subject field while another does not, the latter is excluded from further processing. For example, the "sports" reading of the English noun *goal* will be excluded from syntactic analysis of a text belonging to the subject domain of computing or medicine, because its Russian equivalent, *гол*, is not attested in either of these domains.

English lemma	Russian lemma	frequency	day of the year
activity N _	активность N _	7	55
activity N _	деятельность N _	4	55
address N 1	адрес N _	111	55
address N 1	обращение N 1	4	55
address V 2	обращаться V 1	7	55
allow V _	позволять V _	37	55
allow V _	разрешать V 2	10	55
application N _	приложение N _	11	55
application N _	заявка N _	6	55
call V 2	называть V _	5	55
call V 2	вызывать V 1	6	55
control N 1	контроль N _	7	55
control N 1	управление N 1	10	55
different A _	различный A _	6	55
different A _	разный A _	7	55
directory N _	директория N _	117	55
directory N _	справочник N _	9	55
drive N 2	вождение N _	6	55
drive V 1	водить V _	8	55
feeling N _	ощущение N _	4	55
feeling N _	чувство N _	4	55

Figure 8: Examples of rated translations for the subject domain 'computing'

During the transfer phase, the ETAP-3 system gives priority to those translation equivalents that have been attested for the current subject domain. As the frequencies do not only relate to a word in a subject domain but to the translation hypothesis in a subject domain, we expect this approach to yield better results than the first one. An evaluation which confirms this hypothesis is however still pending.

4 Learning, Forgetting and Remembering

Any systematic collection of data represents a learning process that must be supported by a process of **forgetting**. In fact, as shown by studies in cognitive science, these tasks are so closely linked that forgetting has been claimed to be a vital component of learning (Meh74). So, in language acquisition one learns phonematic oppositions via forgetting unused oppositions.

In accordance with this assumption, we have supplemented the frequency collecting mechanism by a forgetting mechanism. The forgetting mechanism is invoked during every learning session and "attacks" old data with low frequencies by reducing their frequency until they fall out of the active memory. By this procedure, data remain up-to-date and accidental erroneous data are rid of: mistyped words, accidental matches, words coming from other languages or different subject domains are removed.

Naturally, the forgetting mechanism cannot operate on its own. In psychological terms, forgotten or unlearned words, may be learned later more easily than completely new words. Therefore, we allow for the **recollecting** forgotten data. In addition to the so-called active memory which guides the output of the system (e.g. supplies the current frequency list) we implemented a **passive** memory, which stores every occurrence of an

item. Whenever an item is found in the corpus, the updating and counting is performed in this passive memory. In every learning session, the old items with low frequencies (however exceeding the threshold) are identified and their frequencies are reduced by one. At the end of the learning session, all items of the passive memory which exceed the threshold are copied into the active memory, while those below the threshold remain in the passive memory with a frequencies reduced by one. This means that items that have been forgotten are more likely to overcome the threshold in a subsequent session than a completely new item as long as they remain in the passive memory. Whenever a forgotten word is encountered anew, the frequencies stored in the passive memory are updated and copied into the active memory if the frequencies exceed the threshold. If an item drops out of the passive memory, it has to be re-learned completely anew. This process is illustrated in figure 9, where f_q refers to "frequency", f_q++ to "increase frequency by one", f_q-- "reduce the frequency by one" and $THRSH$ to "threshold":

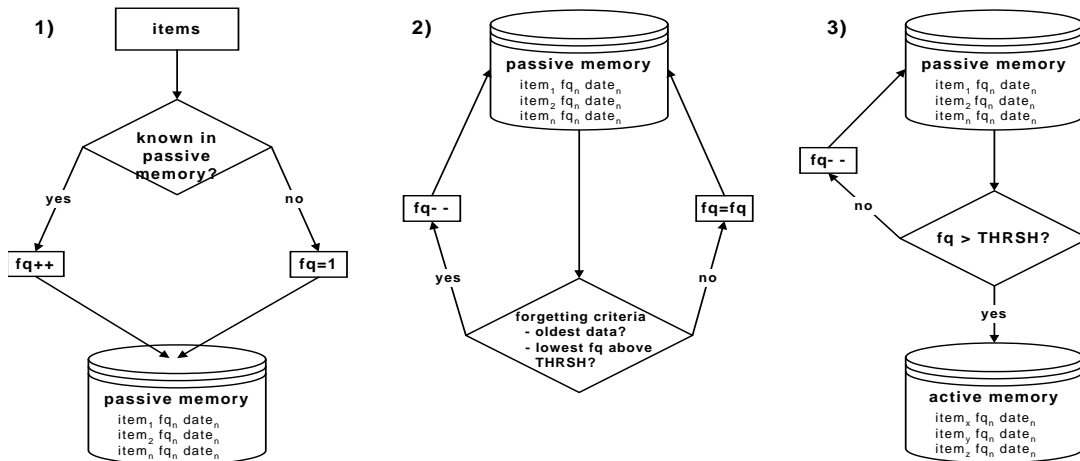


Figure 9: Learning (1), forgetting (2) and remembering (3) of translation frequencies

5 Conclusions

We have shown how monolingual and parallel corpora can be used to support RBMT systems. Word frequencies of target words and frequencies of translation pairs can help to distinguish between words of the source language which otherwise are difficult to distinguish between. The SL analysis becomes more efficient and plausible since non-attested ambiguities are rejected at an early stage.

During the transfer phase of the RBMT system, translation equivalents that have received a high statistical rating in a given subject field are tried first.

As the calculation of frequencies and the compilation of these data into the RBMT lexicon are made fully automatically, adaptation to a new or a more specific subject field can be implemented any time with no human intervention. Consequently, the system is easy to customize, works more efficiently, and produces more reliable translations.

As the RBMT systems we backed with statistical data are the same as those previously linked to memory-based systems, we have actually made an attempt of merging

the advantages of three MT paradigms into one framework. While the memory-based components in this architecture invokes the RBMT component for chunks that it cannot handle, the statistical data help the RBMT component through the analysis and transfer. Chunks coming from the memory-based components are not affected by the statistical component.²

The experiments described above represent a further step towards the integration of different NLP approaches. The potentials of such integration, however, are still far from being explored in full.

References

- L.L. Cinman and L.L. Iomdin. Lexical Functions and Machine Translation. In *Proceedings of the Dialogue'97 International Seminar in Computational Linguistics and Applications*, pages 291–297, Moscow, 1997. (In Russian. English summary).
- Michael Carl, Leonid L. Iomdin, and Oliver Streiter. Towards dynamic linkage of Example-Based and Rule-Based Machine Translation. In *ESSLLI '98 Machine Translation Workshop*, 1998.
- Michael Carl, Catherine Pease, and Oliver Streiter. Examples of hybrid Machine Translation. In *ISMT and CLIP, Beijing*, 1999.
- Shinichi Doi and Kazunori Maraki. Translation ambiguity resolution based on text corpora of source and target language. In *COLING-92*. 1992.
- Robert Frederking and Sergei Nirenburg. Three heads are better than one. In *Proceedings of ANLP-94*, Stuttgart, Germany, 1994.
- L.L. Iomdin and Oliver Streiter. Learning from Parallel Corpora: Experiments in Machine Translation. In *Proceedings of the Dialogue'99 International Seminar in Computational Linguistics and Applications, Tarusa (Russia), June 1-5, 1999*, 1999.
- Jacques Mehler. Apprendre par désapprendre. In *L'unité de l'homme, Le cerveau humain*. Paris, 1974.
- Rita Nübel. End-to-end evaluation in Verbmobil I. In *MT-Summit*, San Diego, 1997.
- H. Nomiyama. Lexical selection mechanism using target language knowledge and its learning ability. In *IPSJ-WG, NL86-8 (in Japanese, cited in (DM92))*. 1991.
- Philip Resnik. Probabilistic tree-adjointing grammar as a framework for statistical natural language processing. In *COLING-92*. 1992.
- Harold L. Somers. "New Paradigms" in MT: The state of play now that the dust has settled. In *ESSLLI '98 Machine Translation Workshop*, 1998.

²The follow-up of the research presented here can be found in (IS99). All papers of the authors cited above may be found at www.iai.uni-sb.de and proling.iitp.ru