Learning from Parallel Corpora: Experiments in Machine Translation

Dialogue'99: Computational Linguistics and its Applications

International Workshop Vol.2, pp. 79-88

Leonid Iomdin, Oliver Streiter

Institute for Information Transmission Problems,
Russian Academy of Sciences
Bol'shoi Karetnyj Pereulok 19, Moscow, 101447, Russia

iomdin@iitp.ru   oliver@rockey.iis.sinica.edu.tw

ABSTRACT The research described in this paper is rooted in the endeavours to dynamically combine different MT approaches in order to improve the performance of MT systems – most importantly, the quality of translation. The authors review the ongoing activities in the field and present a case study, which shows how simple statistical data concerning single- and multiword translations can be drawn from parallel corpora and compiled into the lexicon of a rule-based MT system. As a result, the lexicon is enriched with translation equivalents attested for different subject domains, which facilitates the tuning of the MT system to a specific subject domain and improves the quality and adequacy of translation.

KEYWORDS: machine translation, parallel corpora, hybrid machine translation systems, statistical processing of texts

Обучение на корпусах параллельных текстов: эксперименты по машинному переводу

Л.Л. Иомдин, О. Штрайтер

Институт проблем передачи информации
Российской академии наук
101447, Москва, Большой Каретный переулок 19

iomdin@iitp.ru   oliver@rockey.iis.sinica.edu.tw

РЕФЕРАТ Исследование, описываемое в работе, лежит в русле предпринимаемых в последнее время попыток динамически соединить в одной платформе различные подходы к машинному переводу. Цель такого соединения – оптимизировать работу системы МП и, в первую очередь, улучшить качество перевода. Дается краткий обзор современных работ в этом направлении и излагается серия экспериментов, в ходе которых из параллельного корпуса текстов статистическими методами извлекаются относительно нетривиальные и при этом достаточно частотные переводные эквиваленты отдельных слов и двухсловных сочетаний. Эти эквиваленты автоматически вводятся в двуязычные словари системы МП и корректируются после каждого сеанса статистической обработки, что дает возможность заметно улучшить качество перевода. Эксперименты проводились на системе автоматического перевода ЭТАП-3, разработанной в лаборатории компьютерной лингвистики ИППИ РАН.

1 INTRODUCTION

For decades, the rule-based approach (RBMT) has been the only strategy pursued by researchers in the field of MT. The emergence 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 computer, which accumulates translation knowledge

faster than its human counterpart

in greater quantities

in a format which is directly usable by the machine.

After years of experiments and discussions it has become clear, however, that none of the approaches in their isolated form will solve the problem of MT within a reasonable time (cf. Somers 1998). It is equally unlikely that a new, "ideal" approach may be proposed and implemented on a sizeable scale in the foreseeable future. Substantial progress in the field can therefore be achieved only by combining the strengths of different approaches.

The authors' present attempt to integrate different MT approaches contrasts with earlier static attempts at such combination, in particular, with the experiments in which different translation engines are run in parallel, as in Pangloss (see Brown and Frederking 1995, Brown 1996). Actually, such implementations offer little help, since only the final outputs of different engines are compared. On the other hand, it has already been shown that an integration of different MT approaches may yield better results than those achieved by an individual system. An instructive example is the experience of the Verbmobil platform where the use of the complementary strengths of at least three MT approaches in one framework (deep analysis, shallow dialogue-act based approach and simple TM) improves the performance of the system (see Nubel 1997). What we have tried to do is to supplement an advanced rule-based MT system with a statistics-based preprocessor, which aims at rating translational equivalents offered by the MT system and creating a set of semi-compositional multiword translation equivalents from parallel corpora classified with respect to subject domains.

2 PRIOR RESEARCH

Within the above general research strategy, we tried to find ways of integrating different MT paradigms in one framework and implemented prototypes of such integrated systems. The MT paradigms we considered in earlier experiments were RBMT on the one hand and EBMT or Translation Memories on the other hand.

In one of these experiments, we linked CAT2, an RBMT system (see Streiter 1996, 1998) to an EBMT system EDGAR (Carl 1998). During the analysis phase, EDGAR comes into play after the morphological analysis and before the syntactic analysis performed by CAT2, and, during generation, it operates after the syntactic generation and before the morphological generation. In such an architecture, EDGAR serves for CAT2 as an intelligent multiword and phrase translation front end, whereas CAT2 for EDGAR performs the translation of linguistic structures which are beyond the capabilities of EDGAR.

In a second experiment, we linked the RBMT system ETAP-3 (see Апресян и др. 1992, Apresjan *et al.* 1992, 1993, Иомдин, Цинман 1997) with a TM prototype. During translation, a sentence of the source text is checked against the TM case base. If any of the examples contained there are found, all syntactic links that involve the example concerned are forcefully established prior to regular parsing operation, irrespective of whether the same links would later be obtained or not.

All links that contradict those established for the example are overridden, including the links that originate from those words of the example that are not allowed to have daughters. After the syntactic tree of the sentence processed is fully generated, it is sent to the transfer component, where the equivalent side of the example is substituted for the fragment corresponding to the source side of the example. In the simplest case, the (only) syntactic link coming into the top node of the source example is replaced by a link coming into the top node of the target example, whereas all links originating from the elements of the source example are represented as ones originating from the top node of the target example.

Both experiments in linkage have been recently reported in detail in Carl *et al*. (1999).

In a previous paper (Streiter et al. 1999), we showed how, within an RBMT system, simple statistical data can be used (a) to tune the system automatically to a specific subject domain and (b) to improve the system's efficiency. In the experiment, lexical data were extracted from **monolingual** corpora using statistical methods and compiled into the lexicon of the RBMT system. The enriched RBMT is used during source text analysis to (partially) exclude unlikely homographs and during transfer, where translation equivalents attested for the relevant subject domain are tried first.

A somewhat different approach within the same research paradigm of hybrid MT systems is found in Heyn (1996) and Carl *et al*. (1998b), where the effect of combining TM technology with linguistically rich representations is investigated.

3 BILINGUAL CORPORA SUPPORT FOR MT

The main objective of the present study is to investigate the potentials of integrating statistical data drawn from **parallel bilingual** corpora into an RBMT system. In contrast to recent publications by Choi et al. (1998) and Jung et al. (1998), we did not use statistical information to recognise the subject domain to which the texts belong. Instead, we used corpora classified with respect to the domain to rate the translation hypotheses created by the RBMT system with respect to their likelihood of occurrence in a text from the respective subject domain. The rated hypotheses were transformed into a rule-format and automatically integrated into the MT lexicon. The new data not only proved to be useful for translation but helped to solve some of the ambiguities in the source language which, until now, could not be solved by other means.

3.1 Rating Translation Equivalents

The lexicon of an RBMT system represents, 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 long list of translation equivalents, using several lexica of the ETAP-3 system. The list, called TRANSETAP, which contained over 130,000 bilingual equivalents (for the most part consisting of one word each) supplied with information about the part of speech and, in case of lexical ambiguity, the lexeme number was checked against representative parallel corpora, each associated with a specific subject domain.

3.1.1 Morphological Preprocessing of Parallel Corpora

In order to make such a check possible, every monolingual text of the parallel corpus was morphologically processed. The motivation for the intervention of the morphological analyzer is rather obvious: the reduction of as few as two inflected forms to one underlying lemma renders statistical data more informative. For example, if the source text contains two word forms of the same lexeme, e.g. *player* and *players*, the resulting statistical data will count two occurrences of the lexeme *player*. It is even more important to use information on inflection if the source language has rich morphology, as is the case with Russian, where a verbal paradigm may contain well over 200 word forms.

The morphologically processed texts are transformed into strings of possible lemmas. So, the English sentence *This man mans the new boat* will be eventually presented as

(1) THIS S 1| THIS A 2 || MAN S 1 | MAN V 2 || THE Art _ || NEW A _ || BOAT S 1 | BOAT V 2,

while its Russian counterpart, *Этот человек набирает команду для нового корабля*, will receive the following representation:

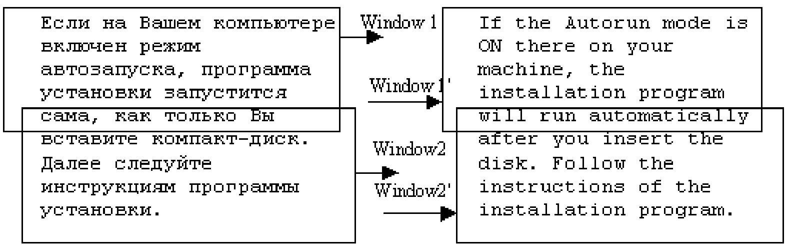
(2) ЭТОТ A _ || ЧЕЛОВЕК S _ || НАБИРАТЬ V _|| КОМАНДА S 1 | КОМАНДА S 2 || НОВЫЙ A _ || КОРАБЛЬ S _

In (1), THIS S 1 and THIS A 2 represent, respectively, the demonstrative pronoun and the pronominal adjective, MAN S 1 and MAN V 2 represent the noun *man* and the verb *to man*, etc. In (2), КОМАНДА S 1 corresponds to the noun meaning *team* and КОМАНДА S 2 to the noun meaning *command.* Homographs are in part automatically reduced with the help of simple contextual rules. Single vertical lines separate different analyses of the same word forms, while double lines separate the words.

3.1.2 Alignment of Parallel Texts

No linguistically sophisticated alignment, either on sentential basis (as in Brown *et al.* 1991) or segment basis identified via anchor points as suggested by Fung (1995) has been attempted. As is known, the former of the two approaches is difficult to implement since parallel texts may be organised differently with respect to their sentence and paragraph boundaries. Texts may also contain pictures, tables, bibliographies etc., which are difficult to align or even text portions present in only one of the parallel texts. The second approach seems unnecessarily complex for the task in question. We have proposed another approach to alignment which, in our opinion, is both simple and adequate for our purposes.

In this approach, the parallel texts, prepared as described in 3.1.1, are cut into partially overlapping parallel strips, or windows (see Fig. 1). The size of the first window is specified by a parameter external to the programme (e.g. 50 words), and that of the second window is dynamically calculated by the programme, depending on the relative size of the two texts constituting the parallel pair. For example, 50 words for the text in the first language may averagely correspond to 48 words in the second text. In this case, the size of the second window will be set at 48 words.



Если на Вашем компьютере
включен режим
автозапуска, программа
установки запустится
сама, как только Вы
вставите компакт-диск.
Далее следуйте
инструкциям программы
установки.

If the Autorun mode is
ON there on your
machine, the
installation program
will run automatically
after you insert the
disk. Follow the
instructions of the
installation program.

Fig. 1. Window alignment of parallel texts

3.1.3 Extraction of Attested Equivalents

The aligned parallel texts are statistically processed by two experimental PERL programmes, run consecutively. The first programme tries to find **multiword translation equivalents** and attests them for the specified subject domain. The second programme, which takes the output of the first programme into account, finds and attests **single-word translation equivalents**. We will start by describing the operation of the programme that deals with single words, because it is simpler, and then proceed to the multiword programme.

The single-word programme compares with TRANSETAP all possible pairs of words taken from two parallel windows (for example, 50 48 combinations for the first window pair, etc.). All pairs {$word_{1m}$, $word_{2m}$} (where $word_{1m}$ belongs to window *m* of Text 1 and $word_{2m}$ belongs to window *m* of Text 2) that could be found in TRANSETAP are stored and counted.

As a result of such processing, a list of **attested translations for a given subject domain** is produced. The following example is an excerpt from the attested translation list generated from a parallel corpus of computer texts. 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 (if any), absolute frequency of the pair, day of the last modification (starting from year's beginning).

activity S _	активность S _	7	55
activity S _	деятельность S _	4	55
address S 1	адрес S _	111	55
address S 1	обращение S 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 S _	приложение S _	11	55
application S _	заявка S _	6	55
call V 2	называть V _	5	55
call V 2	вызывать V 1	6	55
certificate S 1	сертификат S _	67	55
certificate S 1	удостоверение S _	7	55
content S 1	содержание S _	7	55
content S 1	содержимое S _	4	55
control S 1	контроль S _	7	55
control S 1	управление S 1	10	55
design S 2	проект S _	4	55
design S 2	проектирование S _	6	55
different A _	различный A _	6	55
different A _	разный A _	7	55
directory S _	директория S _	117	55
directory S _	справочник S _	9	55
drive S 2	вождение S _	6	55
drive V 1	водить V _	8	55
feeling S _	ощущение S _	4	55
feeling S _	чувство S _	4	55
form S 1	форма S _	46	55
form S 1	образ S _	7	55

As can be seen from these data, the relative occurrence of some of the translation equivalents for the same words can be very different. So, the English noun *address* was translated into Russian as ‘адрес’ 111 times and as ‘обращение’ only 4 times. Similarly, the noun *certificate* was translated as ‘сертификат’ much more frequently than ‘удостоверение’, and the noun *directory* was almost always translated as ‘директория’ rather than ‘справочник’. It is only natural that, for the subject domain concerned, the first equivalents should be preferred in all three cases. The mechanisms ensuring that the attested translation equivalents are produced by the MT system are discussed in Section 3.1.4 below.

With the more complex programme, we try to identify all translation pairs consisting of more than one word that could be produced by concatenating single-word translation hypotheses listed in TRANSETAP, provided they follow one of the general patterns currently covered by ETAP-3 multiword transfer rules, e.g.

Pattern No	**English**	**Russian**

1	Adj + Noun	Y	Adj + Noun
2	Noun1 (comp) + Noun2	Y	Adj + Noun
3	Noun1 (comp) + Noun2	Y	Noun2 + Noun1 (gen)

etc. Thus, if TRANSETAP contains translation hypotheses *personal -> персональный* and *computer -> компьютер*, the translation pair *personal computer -> персональный компьютер* will be identified if the respective word combinations appear in parallel windows, since they match one of the specified patterns (namely, pattern 1 above). The occurrences of multiword translation equivalents are counted, which is used for statistical evaluation in case of concurrent translations. For example, if the programme output yields alternative translation equivalents whose relative frequencies are very different, as in

S:credit1 S:card	A:кредитный_ S:карта_	20
S:credit1 S:card	A:кредитный_ S:карточка_	1

(identified according to pattern 2) it can easily be concluded that the equivalent *credit card -> кредитная карта* is preferable to *credit card ->кредитная карточка* in the subject domain concerned.

Of course, the raw output of such an algorithm, is bound to include, in addition to plain noise, numerous multiword equivalents that are not worth identifying since they are fully compositional, show no variance, and are produced by ETAP-3 without any assistance. For this reason, a relatively simple filtering mechanism is used to delete these combinations from the list of attested translations produced. The mechanism leaves in this list only translations that allow variance (in TRANSETAP and/or parallel corpora) or else are «semi-compositional», e.g. are composed of non-identical sets of parts of speech, as in

S: information S: retrieval -> A: информационный S:поиск

(rather than S:поиск A: информация). Actually, more than 90 % of the entries of this list could be assessed as valuable. An excerpt from the final output of the algorithm, listed below, shows that, even remaining within the semi-compositionality class, interesting and sometimes unexpected equivalents could be found. Many of these are not only absent in the ETAP-3 lexicon but are extremely unlikely to appear in any terminological dictionary - exactly on the ground of their being semi-compositional, i.e. not idiomatic enough. Meanwhile, their usability for an MT system is incontestable. In contrast to the single-word statistical programme which rates the existing translation equivalents, the multiword programme **explicitly creates** some of the equivalents which it then rates. It should also be noted that some of the entries listed below are relevant for both directions of translation while other ones are concerned with one direction only (this happens when an ambiguity of the source language is not accompanied by that of the target language).

A:active_ S:nature_	Y	A:деятельный_ S:натура_	2
A:regional_ S:agency_	Y	A:региональный_ S:орган_	32
A:current2 S:directory_	Y	A:текущий_ S:директория_	1
A:deep_ S:sleep2	Y	A:глубокий_ S:сон_	1
A:diplomatic_ S:relation_	Y	A:дипломатический_ S:отношение_	4
A:due1 S:recognition_	Y	A:должный_ S:признание_	4
A:economic_ S:relation_	Y	A:экономический_ S:отношение_	4
A:elementary_ S:education_	Y	A:начальный_ S:образование_	4
A:friendly1 S:relation_	Y	A:дружественный_ S:отношение_	8
A:fundamental1 S:freedom_	Y	A:основной_ S:свобода_	24
A:general1 S:assembly_	Y	A:генеральный_ S:ассамблея_	689
A:hot_ S:key1	Y	A:горячий_ S:клавиша_	1
A:human1 S:personality_	Y	A:человеческий_ S:личность_	4
A:impartial_ S:tribunal_	Y	A:беспристрастный_ S:суд_	4
A:inaudible_ S:whisper2	Y	A:невнятный_ S:шепот_	1
A:loud1 S:ejaculation_	Y	A:громкий_ S:восклицание_	1
A:methodical_ S:person_	Y	A:методичный_ S:человек_	1
A:particular1 S:circumstance1	Y	A:специфический_ S:обстоятельство_	4
A:personal_ S:certificate1	Y	A:личный_ S:сертификат_	4
A:secret2 S:vote1	Y	A:тайный_ S:голосование_	4
A:sensible_ S:man1	Y	A:здравомыслящий_ S:человек_	1
A:social1 S:insurance_	Y	A:социальный_ S:страхование_	70
S:account1 PR:of_ ART:the_ S:measure_	Y	S:отчет_ PR:о_ S:мера_	1
S:arrear_ PR:in1 ART:the_ S:payment_	Y	S:задолженность_ PR:по_ S:уплата_	1
S:asylum_ PR:from_ S:persecution_	Y	S:убежище_ PR:от_ S:преследование_	4
S:bank_ S:account1	Y	A:банковский_ S:счет_	1
S:carriage_ S:return1	Y	S:возврат_ S:каретка_	1
S:chief2 PR:of_ S:staff1	Y	S:начальник_ S:штаб_	4
S:command_ S:line1	Y	A:командный_ S:строка_	1

S:configuration_ S:file1	Y	S:файл_ S:конфигурация_	4
S:family_ S:planning_	Y	S:планирование_ S:семья_	3
S:human2 S:right3	Y	S:право_ S:человек_	78
S:exchange1 S:rate1	Y	A:обменный_ S:курс_	4
S:land1 S:force2	Y	A:сухопутный_ S:сила_	8
S:information_ S:product_	Y	A:информационный_ S:продукт_	6
S:information_ S:system_	Y	A:информационный_ S:система_	7
S:maintenance_ PR:of_ S:peace_	Y	S:поддержание_ S:мир_	4
S:means_ PR:of_ S:communication_	Y	S:средство_ S:сообщение_	4
S:plan1 PR:of_ S:action_	Y	S:план_ S:действие_	3
S:production_ S:reduction_	Y	S:сокращение_ S:производство_	2
S:rate1 PR:of_ S:decrease1	Y	S:темп_ S:снижение_	2
S:reduction_ PR:in1 S:volume_	Y	S:сокращение_ S:объем_	2
S:right3 PR:of_ S:man1	Y	S:право_ S:человек_	8
S:root1 S:directory_	Y	A:корневой_ S:директория_	3
S:sale1 S:market1	Y	S:рынок_ S:сбыт_	2
S:trusteeship_ S:agreement_	Y	S:соглашение_ PR:о_ S:опека_	28
S:trusteeship_ S:agreement_	Y	S:соглашение_ PR:по_ S:опека_	2
S:world_ S:leader_	Y	A:мировой_ S:лидер_	3
S:world_ S:price1	Y	A:мировой_ S:цена_	2
S:world_ S:war_	Y	A:мировой_ S:война_	4
V:adopt_ S:declaration_	Y	V:принимать_ S:декларация_	3
V:adopt_ S:resolution_	Y	V:принимать_ S:резолюция_	3
V:enter_ PR:into_ S:force2	Y	V:вступать_ PR:в_ S:сила_	6
V:follow_ S:principle_	Y	V:следовать_ S:принцип_	4
V:follow_ S:scenario_	Y	V:следовать_ S:сценарий_	1
V:gain2 S:access1	Y	V:получать_ S:доступ_	1
V:have_ S:access1	Y	V:иметь_ S:доступ_	1
V:make1 S:recommendation_	Y	V:делать_ S:рекомендация_	32
V:obtain_ S:permission_	Y	V:получать_ S:разрешение_	1
V:oil2 S:industry_	Y	A:нефтяной_ S:промышленность_	4
V:take_ S:measure_	Y	V:принимать_ S:мера_	4
V:window2 S:manager_	Y	A:оконный_ S:менеджер_	2

At the end of operation, this program marks in the parallel corpus all occurrences of the translation equivalents found to exclude them from the output of the single-word statistical programme, which is run immediately after it. The reason for this exclusion is that the attested multiword equivalents, to our opinion, must override the single-word equivalents in case of any discrepancy. Accordingly, multiword equivalents, whose frequency of occurrence may be rather high, should not distort the statistics for single-word translations. Indeed, irrespective of what frequency distribution is obtained for the translational equivalents of the noun *man* (e.g. *человек* vs. *мужчина*), the expression *rights of man* must be translated *права человека* and not as *права мужчины* - in full conformity with the multiword translation list given above.

3.1.4 Incorporating Attested Equivalents into the Lexicon and Using Them in Transfer

The translational equivalents produced by the two programmes are automatically introduced into the ETAP-3 lexicon. The transfer mechanism of the system should be instructed to use these equivalents by default and thus override any existing equivalents (whatever they are). In other cases, where the order of magnitude of the relative occurrence of the equivalents is the same (cf. e.g. *application* 'приложение'/'заявка', *feeling* 'ощущение'/'чувство' etc.) as quoted in the first list, both equivalents should be introduced into the lexicon with an appropriate domain attestation tag.

To be more specific, the entries of combinatorial dictionaries of the ETAP-3 system contain different translation fields for different subject domains. These can be created, deleted or modified automatically in accordance with the list of attested translations obtained. For example, the Russian translation fields for the nouns *directory* and *application* will be supplemented by the following records

DIRECTORY

....

DOMAIN: COMPSCI-DOMAIN statistically produced!

TRANS: ДИРЕКТОРИЯ

…………

APPLICATION

…………

DOMAIN: COMPSCI-DOMAIN statistically produced!

TRANS: ПРИЛОЖЕНИЕ/ЗАЯВКА

…………

The above entries should be understood as follows: The word DIRECTORY is TRANSlated for the DOMAIN of COMPuter SCIence as ДИРЕКТОРИЯ; the word APPLICATION is translated within the same subject domain as ПРИЛОЖЕНИЕ or ЗАЯВКА, whereby ПРИЛОЖЕНИЕ is used as the first choice.

As for multiword translations, they are incorporated in the ETAP-3 dictionaries as instantiations of some of the transfer patterns - also mentioning the subject domain. For reasons of space, examples of the respective entry records are left out.

During the transfer phase, the ETAP-3 gives, in the absence of contextual constraints which influence the translation, priority to those translation equivalents that have been attested for the current subject domain and, accordingly, entered into the dictionary.

With this relatively simple measure the translation quality for specific subject domains can be noticeably improved. Likewise, the MT system can be automatically tuned to different subject domains (and even user groups). The only prerequisite for such tuning is the presence of translation examples in the form of parallel corpora.

3.2 Source Text Analysis Assisted by Attested Translations

The rating of translation equivalents described above can also be used to reduce the ambiguity of the source text – and, consequently, processing time - at an early stage of ETAP-3 operation. This is achieved as follows. If the syntactic analysis of a sentence identifies a lexical ambiguity so that (at least) one of the alternative word hypotheses bears a translation field attested for the current subject field while another does not bear such a field, 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, thus reducing the set of parsing variants, because its Russian equivalent, *гол*, is not attested in either of these domains.

3.3 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 (cf. Lorenz 1973). 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 latter, invoked during every learning session, "attacks" two types of data: old data with low frequencies and the data whose frequencies are below a threshold value (which can be specified as required).

The forgetting process serves three main purposes.

First, it ensures that the data remain up-to-date. Secondly, much of the erroneous data is rid of: mistyped words, accidental matches, words coming from other languages or different subject domains are removed. Finally, the forgetting mechanism functions as the evaluation monitor for the set of data. Thus, we do not evaluate them by classical statistical calculi but use the forgetting mechanism to transform the fuzzy data extracted from the corpora into categorial patterns.

Naturally, the forgetting mechanism cannot operate on its own. A rule of thumb says that half of any text consists of words having the frequency 1, independent of the text size. A forgetting mechanism working on words with a low frequency necessarily makes mistakes and is bound to obliterate «good» words, e.g. words which are typical for the text and worth retaining. If we allow the forgetting mechanism to simply remove all these words, we will have to wait until these words reappear at least once during the learning session with a frequency above the threshold value.

In order to solve this dilemma, we have supplied the mechanism with a means of **recollecting** the 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 done in this passive memory. In every learning session, the oldest item(s) with the lowest frequencies (however exceeding the threshold) are identified and their frequencies are reduced by 1. This means that items that have been forgotten are more likely to overcome the threshold in a subsequent session than a completely new item. Whenever a forgotten word is encountered anew, it is treated as an unforgotten word. The frequencies stored in the passive memory are updated and copied into the active memory if the frequencies exceed the threshold, otherwise they remain in the passive memory and are also reduced by 1. This process is illustrated in Fig. 2.

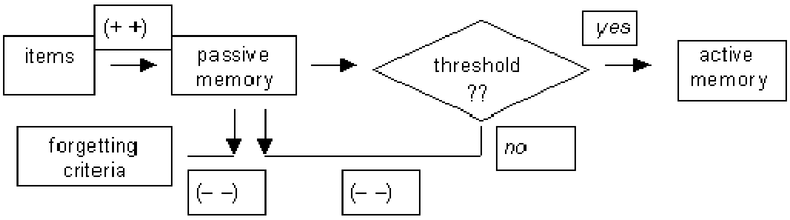


Fig 2. Movement of Statistical Data on Parallel Corpora

4 CONCLUSIONS

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

During the transfer phase, translation equivalents that have received a high statistical rating in a given subject field are used as first choice or default.

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

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

REFERENCES

Apresjan, Ju.D., Igor M. Boguslavskij, I.M., Iomdin, L.L., Lazurskij A.V., Sannikov, V.Z. and Tsinman L.L. (1992). The linguistics of a Machine Translation System. *Meta*, **37** (1): 97-112.

Apresjan, Ju.D., Igor M. Boguslavskij, I.M., Iomdin, L.L., Lazurskij A.V., Sannikov, V.Z. and Tsinman L.L. (1993). Systeme de traduction automatique {ETAP}. *La Traductique.* P.Bouillon and A.Clas (eds). Les Presses de l'Universite de Montreal, Montreal.

Brown, D.R. (1996) Example-Based Machine Translation in the Pangloss System. *COLING-96. The 16th International Conference on Computational Linguistics. Proceedings.*

Brown, D.R., Frederking, R. (1995) Applying Statistical English Language Modelling to Symbolic Machine Translation. *Theoretical and Methodological Issues in Machine Translation..*

Brown, P.F., Lai J.C., and Mercer, R.L. (1991) Aligning Sentences in Parallel Corpora. *Proceedings of the 29th Annual Meeting of the ACL.*

Carl M. (1998) A constructivist approach to Machine Translation. *Proceedings of NeMLaP3/CoNLL98*. Sydney. 247-256.

Carl, M., Iomdin, L.L., Streiter, O. (1998a) Towards dynamic linkage of example-based and rule-based machine translation. *ESSLLI '98 Machine Translation Workshop,* Saarbrucken.

Carl, M., Schaible, J., Pease, C. (1998b) Enhancing translation memory (TM) technologies with linguistic intelligence. *MULTI-DOC Deliverable D4.1.* Commission of the European Communities, Luxembourg.

Carl, M., Iomdin, L.L., Pease, C., Streiter, O. (1999) Towards a dynamic linkage of example-based and rule-based machine translation. *Machine Translation* (to appear).

Choi, Sung-Kwon; Jung, Han-Min; Sim, Chul-Min; Kim, Taewan; Park, Dong-I; Park, Jun-Sik; and Choi, Key-Sun. (1998) Hybrid approaches to improvement of translation quality in web-based English-Korean machine translation. *COLING-ACL-98. The 17th International Conference on Computational Linguistics. Proceedings.*

Fung, P. (1995) A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. *33rd Annual Meeting of the Association for Computational Linguistics*.

Jung, Han-Min; Yuh, Sanhwa; Sim, Chul-Min; Kim, Taewan; and Park, Dong-I. (1998) A domain identifier using domain keywords from balanced web documents. *First International Conference on Language Resources & Evaluation.* Granada.

Heyn, M. (1996) Integrating Machine Translation into Translation Memory Systems. *European Association for Machine Translation - Workshop Proceedings.* ISSCO, Geneva. 111-123.

Lorenz, Konrad. (1973) Die Rьckseite des Spiegels. Versuch einer Naturgeschichte menschlichen Erkennens. R. Piper & Co. Verlag. Munchen.

Nubel, R. (1997). End-to-end evaluation in Verbmobil I. *MT Summit*. San Diego.

Somers, H. L. (1998) "New Paradigms" in MT: The state of play now that the dust has settled. ESSLLI '98 Machine Translation Workshop, Saarbrucken.

Streiter, O. (1996) Linguistic Modeling for Multilingual Machine Translation. Berichte aus der Informatik Series. Shaker Verlag, Aachen.

Streiter, O. (1998) A semantic description language for multilingual NLP. Papers presented at the Tuscan Word Centre - Institut fur Deutsche Sprache Workshop on Multilingual Lexical Semantics, 19-21 June 1998

Streiter, O., Iomdin L.L, Munpyo Hong, and Ute Hauck. *Learning, forgetting and remembering: Statistical support for rule-based MT.* Paper submitted to TMI'99 conference.

Апресян, Ю.Д., Богуславский И.М., Иомдин, Л.Л., Лазурский, А.В., Митюшин, Л.Г., Санников, В.З., Цинман, Л.Л. (1992) Лингвистический процессор для сложных информационных систем. Наука, Москва. 256 с. (In Russian).

Иомдин, Л.Л., Цинман, Л.Л. (1997) Лексические функции и машинный перевод. *Труды Международного семинара Диалог'97 по компьютерной лингвистике и ее приложениям. Ясная Поляна, 10-15 июня 1997 г.* Москва. 291-297 (In Russian. English Summary).