International Congress University - Vilnius 2005

A new way to machine translation

17h00-18h00 mar IKU8 (Witkam), Lapenna

Toon Witkam

Born 1944. Studied aeronautics at the Technical University in Delft (Netherlands). It followed a career in the software industry, of which he devoted most of his time to machine linguistics. At the Dutch company BSO in 1979 he initiated the research project DLT (Distributed Language Translation) for automatic translation with Esperanto as an interlanguage. In 1990 he worked as guest researcher at ATR Interpreting Telephony Research Laboratories in Kyoto. In 1991 he advised the European Commission on machine translation, and from 1992 to 1996 he was professor of computer science and cognitive ergonomics, again in Delft. Since 2002, he has been involved in word statistical analysis of Esperanto texts.

Summary

After keen research projects on machine translation have blossomed twice in the past century, the ambitious goal of high-quality machine translations has not yet been realized. Today, everyone an internet user can immediately get a machine translation from several languages, sometimes even free, but never infallible. In fact, the current commercially available translation technology still rooted in the early development period of the 1960s and 1970s.

However, this does not mean that research efforts have stopped. The factor of automatic translation is still very much alive, and during the last ca. A firm new paradigm has been developed for 15 years: Statistical Machine Translation (PAT). Contrary to previous research, the linguistics that culminating in the 1980s, PAT is surprisingly based on the craft of the human translator himself. Unlike the linguist, the translator processes texts much more practically, usually from the point of view of intercultural communication and not according to the abstractions of linguistic theory.

Precisely because of this, for a long time the translator's experience seemed less 'formalizable' than the linguist's knowledge. But in the end the concepts of the research changed: the linguistic knowledge did not yield sufficient results, and the highly advanced memory capacity of computers in the meantime makes it possible to use huge ones corpora of human translations as a basis for example.

Page 2

Just as the craftsmen in translation bureaus continue to learn from the previous own work and that of more experienced colleagues, computers could do that too. It is necessary that the computerized 'translation memory' should be reliable, extensive, and continuously updated. Technology, the challenge exists in automatically translating by analogy those translators or phrases that are not literally can be found in the example base. Several hundred PAT researchers around the world are concentrating now to that end, and it will take a few more years for the technology to mature.

It is worth mentioning that Esperanto is the work of Victor Sadler (1989, as part of a DLT project) is recognized in the factual field as one of the first PAT investigations. However, so that Esperanto can to really take part in this new path to machine translation, you definitely need a lot of current human translations into our language.

A new road to Machine Translation

After repeated research efforts on machine translation in the 20th century, the ambitious goal of fully automatic high-quality translation has not been realized. Today, internet users can get computerized translations from various languages instantly, even for free, but far from flawless. In fact, the translation technology now on the market is still rooted in the pioneering developments of the 1960s and 1970s.

However, research efforts have not come to a halt. Machine Translation is still an active field of R&D, and over the last 15 years a new paradigm evolved: Statistical Machine Translation (SMT). In contrast to research based on linguistics, which culminated in the 1980s, SMT is surprisingly based on the craft of human translation itself. Unlike the linguist, the translator treats a text in a much more practical way, usually with communication-between-two-cultures in mind, rather than the abstractions of theoretical linguistics.

For this very reason, translator's skill always appeared much more difficult to 'formalize' than linguist's knowledge. But finally, the research community took a turn: linguistics did not deliver enough results, and the meanwhile dramatically increased computer capacity enabled extremely large corpuses of human-translation examples.

Just like humans acquire skill by learning from previous work or experienced colleagues, computers might act similarly. Of course, the computer's `translation memory 'must be reliable, inexhaustible and constantly updated. Technologically, the challenge exists in

automatic translation-by-analogy of all phrasings not literally found in the example database. Worldwide, several hundred SMT researchers now concentrate on this and it will take some years before the technology ripens.

As to Esperanto, the work of Victor Sadler (1989, as part of the DLT project) is being recognized as one of the earliest SMT researches.

Page 3

However, in order to really profit from this new prospect for machine translation, the Esperanto community urgently needs to produce a vast amount of up-to-date human translations into its language.

A new approach to machine translation.

Research in the 20th century on machine translation has not reached the expected results. Today internet users can instantly get one Automatic translation from various languages, sometimes even for free, but never without errors. In fact the translations available in the market are always made according to the first technological developments in this field dating back to the 60s and 70s.

However, the search did not stop there. It continues in the field, and during the In the last fifteen years a new paradigm has developed: Machine Translation Statistics (TAS) **)** . Unlike research that relied on linguistics and that has known at its heyday in the 1980s, *CAS is based* on the art of the translator. Unlike the linguist, the translator approaches the texts in a more pragmatic way, emphasizing the “communication between two cultures” aspect, rather than linguistic theory.

It is precisely for this reason that the art of the translator has long seemed more difficult to formalize that linguistic theory. But eventually the search changed direction. La linguistic did not give satisfactory results, and thanks to the phenomenal increase of the memory capacity of computers, huge corpora of examples of manual translation as a database.

Just as translators in their office are progressing through their previous work and that of more experienced colleagues, the same goes for computers. It is necessary of course the translated examples entered into the computer are reliable, abundant and constantly updated. Technologically, the challenge is to translate automatically by analogy all expressions or phrases that do not appear in the database. Hundreds of researchers around the world are focusing on this problem and it will take a few more years for the technology to be developed.

Regarding Esperanto, Victor Sadler (1989, as part of his work for the projet *DLT*) est considéré comme l'un des pionniers of the *TAS* . However, to be able to really enjoy this new breakthrough in machine translation, the Esperanto-speaking community must build a vast corpus of human translations in this language.

Page 4

International Congress University - Vilnius 2005

A new path to Machine Translation

Toon Witkam
Utrecht, the Netherlands
[toon.witkam@planet.nl]

There is a large machine translation center in Hyderabad. There, a battery of computers
at night I read a lot of newspapers, magazines, books, reports, from all over the world. Mostly by
internet, occasionally by scanning paper archives. Those indefatigable computers don't do that for
pleasure. Their human masters trained them for that relentless reading. In fact, the training has already continued
decades, and has been perfected to the extent that the brave machines are only daily maintained and updated
their translation talent through a lot of reading and comparison of multilingual text sources, dealing with
the same subject. The dream of an ambitious translator!

Twenty years ago, the grandparents of these hajdarabad machines were still beginners on the
translation card, when each day of comparative reading meant a modest improvement on a wide scale of
translation expert. But the electronic sons and finally the grandchildren, once again faster and more memorable, got it
a level of quality higher than that of the average human translator.

The hajdarabada center therefore serves to support that quality level - a level that is fast
will go down if the translation machinery does not adapt to the continuous changes in the world: new
terms, new languages, new acronyms, and for political reasons new geographies, and even
reviving languages. Precisely for this reason, an international team of professionals - polyglots, media
specialists, scientists - guard their machine classmates. They check if the devices are not
ignores some sources, nor abuses others; whether they are not subtly infected, or their translation
remains reliable, etc.

Except in India, there are similar large translation machine power plants in the United States (San Diego),
Europe (Nancy) and Korea (Pyongyang). All four use different machines and software,
therefore the four operate independently of each other. Just like in the past experienced humans
translators did not deliver an exactly identical translation of the same source text, nor did the four exchanges
does that.
This happy plurality is skilfully exploited by online translation services. They have software that a lot
precisely checks the concordance of the four translations. With this, they can guarantee to their clientele
the more reliable a high-quality translation product.

Here is a projection of the state of machine translation in the year 2055. Is it science fiction? Not really. La a fictitious element lies more in the organizational or corporate side of the sketched power plants than in the technology itself. Who will be willing to invest widely in the issue? Business companies are looking to play robots as more attractive to the market. Governments, multinational institutions?

Page 5

Technology on our plate

In today's world of 2005, lightning-fast machine translation (AT [16]) is already available: per internet, and often even for free. The fact that such translations are not of high quality is hardly important for majority of its users. In the meantime, people have become accustomed to inaccurate, negligent language. use in general, so the less shocking a slightly flawed translation appears.

On the other hand, where it comes to texts to be published, whether public service communications or manuals in industry, careful writing of the original - as far as possible according to a certain standard or model in question vocabulary and sentence construction - can create the right conditions to apply AT technology high quality. Examples of this are the Canadian translation system Météo for weather forecasts, the French and US multilingual Titus and Caterpillar systems for the textile industry and export of agricultural machines, special systems for multilingual software documentation, etc. There is also consulting firms that adapt an existing AT system like Systran to the needs of international companies. In all these cases, it is a question of a high-quality translation of a language fraction [17] .

Except for such for specific application and customer tailored systems, the AT quality of publicly accessible translation services on the web are still very modest. Not just syntactically and stylistically, either regarding the correct word choice. One cannot even guarantee that today's machine translations are reliable conveys the meaning of the original. After a recent survey [Hutchins 2003], recognized a specialist in AT history concluded that the translation quality [18] has barely grown in the last two decades.

Note that *the new road* this paper is about is still an exploration road! Its emergence until now has not reached the AT technology ready to use on the market. This is actually based on research- and development work from the sixth to ninth decades of the last century.

The lack of impressive results, after decades of research and development, in itself was sufficient a challenge for next-generation scientists. New generation - not only in age, but rather in discipline and methodology. The professionals who worked at AT during the 1950-1990 era were computer-scientists, polyglots, grammarians, lexicographers, semantics, logicians, practitioners of the formal linguistics, but hardly professional translators. The biggest project ever, EUROTRA, eventually failed due to an excess of theoretical linguists [Hacken 2001].

Learning machines instead of linguists

The new generation of AT researchers emerged around 1990. It has grown over the last 15 years and

gradually replaced the previous one. But since its birth [Brown 1988], it has entered a whole new path. Giving up any linguistic knowledge, the newborn researchers 19 turned to the model of professional *human translator* ! This is almost a revolution, certainly a paradigm shift. Do one now research experienced translators, keep an eye on them while working, guess the processes in their brains?

More simply: study only the results of their work, the translated texts, of course with the originals nearby. Or more conveniently: let computers do that. Great idea! Computers abounds, becomes faster and less expensive every year, has a huge memory, and doesn't complain about

16 Throughout the text, we will use the abbreviation AT (Machine Translation), in accordance with MT (*Machine Translation*) in English language texts.
17 in English *'controlled language'* or *'sublanguage'* .
18 Hutchins tested AT systems like 'Systran Personal', 'Personal Translator' and so on. and online AT services like Bablefish, Lycos, Reverse, Promt, FreeTranslation, InterTran.
19 instead of linguists mainly mathematicians and computer scientists.

Page 6

long working time. Moreover, there was indeed the concept of a 'learning machine' 20 , whose history goes back to the 1950s, so more or less as far as that of machine translation, though the two lands had not barren each other.

So since approx. 1990, the history of AT, on its very new path, is certainly about the evolution of learning machine. In fact, it concerns both development and development, because the issue also depends on good teachers! Here are the general principles of computer-assisted translation learning:

➢ HUGE QUANTITY of learning material. The learning computer has to read through a lot examples of good (human) translation: at least tens of thousands, preferably millions of sentences from so-called bilingual or multilingual corpora. Here are a few:

- Hansard (Canadian parliamentary debates):
  English and French, 2,000,000 sentences each, 40,000,000 words each
- EuroParl (EU parliamentary debates, 1996-2001):
  11 languages, 740,000 sentences each, 20,000,000 words each
- CNS (Chinese Internet News Service):
  English and Chinese, 25,000 sentences each; 500,000 English words; sentence pairs eg:

➢ CONTINUOUS COUNTING. To be explained here is the above-mentioned act of 'reading through'. Means that the machine systematically traverses each sentence pair or sentence multiple, recording and counting its elements (words, word sequences,…) and certain relationships between them, all according to the precise instructions of the teacher. The counting feeds a probability calculation, by which the machine will later (in the test phase) itself try to translate unseen sentences.

➢ TEACHING TO MAKE PEOPLE. The researchers are the teachers. Maybe no longer necessary in 2055, but now still the protagonists. Every AT researcher, or at least everyone research team, teaches the machine in its own way, experimenting with its own method or method-variation. Some researchers give very simple instructions to the student computers, others teach them in more detail or in a complex way. In most cases, the various methods are

published, and there are international departmental meetings annually. Much of this can be found also on the Internet.

Just outlined is the LEARNING PHASE. The corpus used in this phase - or the part used for learning of it - is the LEARNING BODY 21 . As soon as the machine, like any student, has to show its knowledge, begins the TEST PHASE, which will one day eventually be transformed into a PRODUCTION PHASE.

In the test phase, sentences from the same corpus as the learning corpus are presented to the machine. That is fair, because a corpus represents a certain type of text whose learning the test must control. Pro that is, the corpus used is always divided into two parts: the corpus, usually the largest part, and a small subset of test sentences. Of course, the student should not see the test sentences while learning!

20 English: *'machine learning'* .
21 English: *'training corpus'* .

Page 7

The repeated testing of machine translation machines, a characteristic of the new way, was urgent introduction of a more or less automatic outcome assessment. The last five years, some word statistics aids to measure the quality of AT products are invented 22 , but their use is still disputed. Moreover, in addition to the quality measure, the type of errors must also be studied.

The really fastest phase is - counter-intuitively - the learning phase, despite the sheer amount of the learning material to be worked on. The learning phase is indeed fully automatic[23], and computers become every year faster.
In other words, the fastest and most expensive phase is the TEACH PREPARATION phase: the human work of researchers who continue to invent new teachings for their patient machine adherents.
This phase is also the most creative part of the cycle.

Here are some key points of the issue. There are further accessory phases, e.g. the annoying and temporaba previous one phase to clean and prepare 24 text corpora, to install and configure the various software, to phrase the corpus, etc.

Ungrammatical pirates are pioneers

The pioneers of the new path were mathematicians in a research center of the American company IBM 25 . Them inspired previous advances in the technology of automatic speech recognition - until then a whole other research field than text translation. Noting that *"The problem of language modeling for AT basically equals that for speech recognition"* [Brown 1993], they developed a solid mathematical basis for pure Statistical Machine Translation (PAT 26). Their 1993 magazine article, with 20 full pages of mathematical formulas, became the most referenced source in the AT field. The one proclaimed in it Fundamental Equation of Machine Translation:

$$\hat{s} = \operatorname{argmax}_{s} \Pr(s) \Pr(t \mid s)$$

summarizes the three challenges of PAT: computationally estimating the probability of a source language model Pr (s), computationally evaluate the probability of a translation model Pr (t | s), and invent an efficient, yield a way of searching to find that source language string that maximizes the product of those two probabilities 27 .

The language model is only about the linear sequence of words in a sentence, without any use of syntactic knowledge. In the translation model known as IBM Model 1 word sequence is even completely missing:

instead of words from left to right on a line, they were thrown into a bag for sentence lengths.
Word position does not exist there, and the fact that by language model of word triples 28 (taken over by speech recognition system) 84 percent of test sentences 29 proved to be automatically reconstructible from theirs word bags, indicates the strength of simple word statistics.

To gain some intuitive understanding of PAT, imagine yourself in the role of the (learning) machine. Oni confronts you with hundreds of thousands of pairs of sentences, in two languages completely unknown to you. Nor grammars nor dictionaries are available. In addition, to make the exercise more fun, everyone was prepackaged

22 ia BLEU *(BiLingual Evaluation Understudy)* , NIST *(National Institute of Standards and Technology)* , RED *(Ranker based on Edit Distances)* , ORANGE *(Oracle Ranking for Gisting Evaluation)* .
23 with the exception of a few, such as the "Linear B" PAT system, built in Edinburgh [Callison-Burch 2004].
24 English: *'preprocessing'* .
25 IBM Research Laboratories, Yorktown Heights, NY, USA.
26 Throughout the text, we will use the abbreviation PAT (Statistical Machine Translation), as in English text SMT equals *'Statistical Machine Translation'* .
27 the letters s and t means vortĉenon in respectively the source language (*'source language'*) and the target language (*'target language'*); ŝ means that a source language string that maximizes the probability Pr (s | t), i.e. the probability that a (reverse) translation of a string t will produce the string s.
28 in English *'word trigrams'* or simply *'trigrams'* .
29 were sentences with a maximum length of 10 words [Brown 1990].

Page 8

a sentence in a bag, so that its words 30 are completely disordered. So you now have hundreds of thousands before you of sacks. You will have years of time (by comparison, for a modern computer machine, a millisecond indeed lasts a year), so you bravely embark.
You open the two bags of a first pair and blindly take a word from both. From the bag of Language-1 appears the word *'krht'* , from that of Language-2 the word *'uaaio'* . You reason: "Interesting! The sentence in one bag is the translation of the one in another bag, so there is some chance that the word *'uaaio'* is a translation of the word *'krht'* ".
You are now searching all the pairs for the presence of those two words. Obviously you count the times that both words appear in the same pair, but you also count the one-sided appearances. It matters the relative frequency. If *'krht'* appears in almost every sentence, it is probably a common function word (such as 'and' in Esperanto). If *'uaaio' is also* present everywhere, even then the frequent presence of the two words do not imply that they are the translation of each other (*'uaaio'* could be another common function word). The ideal case is the finding that *'krht'* appears only in e.g. promilo da sacks, always only at the same time as *'uaaio'* , and vice versa. This would result in: Pr (*'uaaio'* | *'krht'*) = 1. More likely, *'krht'* and *'uaaio'* will also appear sometimes without each other, or even several times if one of the two words is of the polysemous type. Then their translation probability would be calculated at e.g. 0.95 or only 0.65.

Thus, in the first experiment of the IBM research group [Brown 1990], the machine computed for each combination of 9000 English and 9000 French words 31 the probability that it is a translation pair, which therefore resulted in an array of 81,000,000 parameters. Its value is: to indicate temporarily word alignment 32 , an important concept in PAT, aptly illustrated by lines between two sentences linking those words that are a translation of each other.

However, the process outlined above was just the initial step in a row of successors: the IBM Models 2-5 (invented all around 1990), passed iteratively, make the previously calculated probabilities more precise. parameters, based on the following information: word position, fertility 33 , and distortion 34 .
The role of word position is obvious: translations of words from the initial part of the source language-- a sentence will probably also appear in the initial part of the target language sentence, etc. A beautiful example of word alignment completely corresponding to word position is 35 :

Among the many questions raised by the expanded membership of the European Union is the question of languages.

Among the many questions raised by the enlarged membership of the European Union is the question of languages.

In this sentence pair, the regularity of positional alignment of word translations is exceptionally fortunate! In most cases, sentence translation has one or two alignment distortions, for example due to the inversion of the adjective-noun-sequence or due to differences in SVO (Subject-Verb-Object) word order (*'je le voit'* - 'I see him'), when translating from English (or Esperanto) to French. But between e.g. the Japanese and the English, the differences in word order are more persistent. Therefore, if the word position information does not really contribute, today researcher uses only the IBM model 1, not the IBM models 2-5 [Ding 2003].

A ubiquitous phenomenon is the so-called 'fertility'. When a source language word produces two- instead of one-a word translation in a target language sentence, its fertility equals 2 instead of 1. A prominent example

30 the number of words in a bag (sentence) can vary between about 10 and 30.
31 the computation was limited to the 9000 most common words that appeared in the corpus.
32 English *word 'alignment'* .
33 English *'feritility'* .
34 English *'distortion'* .
35 sentence pairs from the (Esperanto and English) fact sheets about the Fourth Nitobe Symposium (Vilnius, 30 July - 1 August 2005).

Page 9

is the English function word *'not'* translated into French *'ne... pas'* . But also with content words, fertility> 1 abounds. See the following sentences 36 :

Tensions | between | the | two | powers | have increased | in | recent | months.

The tension | between | la | du | great regional powers | grew | while | the last | months.

If we consider the Esperanto sentence as a source language, the word 'grew' has a fertility of 2, because it produces *'have increased'* . Otherwise, the English words *'tensions'* and *'recent'* have that fertility, and the word *'powers'* has even fertility 3: 'great regional powers'. Three probably, in the same body of text there are also pairs of sentences in which *'powers' are* lined up differently, e.g. with 'great powers', simply with 'powers' or with 'power'. The essence of PAT is that it captures in its probability parameters all the variations found in a bilingual text corpus, therefore in fact the products of the translators' experience and freedom - not the rules of grammar or the dictionary information. This is exactly where PAT, the new way, differs from the traditional AT.

IBM's ungrammatical pioneers, after a per-corpus 37 learning phase of 40,000 sentence pairs French, with a total of approx. 1,600,000 text words, achieved the following result [Brown 1990]: their a learning machine was able to translate 48 percent of 73 French test sentences well. Modest success, which was nevertheless encouraging and inspiring. Certainly impressive was their second experiment [Brown 1993], in which the machine had 1,778,620 sentence pairs, calculated the translation probabilities of 2,437,020,096 word combinations, and using a purely statistical algorithm computed the correct word alignment from for example the 1.9×10^{25} theoretically possible word alignments in the following sentence pair:

What is the anticipated cost of administering and collecting fees under the new proposal?

Under the new proposals, what is the expected cost of administration and collection rights?

Finally, the merit of IBM's PAT pioneers in the early 1990s is also that they clearly agreed to a subsequent necessary addition of morphological and syntactic components to PAT. The size value of their work remains: they have appropriately introduced statistical methods into the AT field, and

convincingly showed its strength.

Syntax silently returns

In sync with the development at IBM in the early 1990s in the United States, but regardless of it, a new paradigm was developed in Japan. It was close to PAT, but retained the syntax: Example-Based Machine Translation (EBAT 38). A common feature of PAT and EBAT is the guidance on texts from the translation practice, through the use of a bilingual corpus or database. Like the IBM researchers, as well as Japanese peers, were partly inspired by computer work speech recognition.

The first prototypes of EBAT were made [Sato 1991]. He initially experimented with bilingualism database of example phrases. Here is a snapshot of a spreadsheet (with the phrase words re-ordered according to VSO 39) used inside its prototype system:

36 source: "China vs. China", an article by Ignacio Ramonet in *Le Monde Diplomatique* of April 2005, translated into English by Ed Emery and translated into Esperanto by Wilhelm Luthermann; the English and Esperanto texts, although both are translations, can themselves be used as (part de) English-Esperanto corpus in PAT research.
37 the *Hansard* Bilingual Corps , an archive of Canadian parliamentary debates.
38 in English-language facts: EBMT *('Example-Based Machine Translation')* .
39 VSO = Verb-Subject-Object.

Translating a new phrase (*'japanese play card'*) from English into Japanese means counting it semantic 'distance' to each example phrase with the same verb (*'play'*). The calculation works with a Japanese-language thesaurus 40 , to which English words have also been added. Thus the machine finds the example closest to the phrase to be translated, and can translate that accordingly.

Makoto Nagao, the master of Japanese AT researchers, who launched the idea of EBAT already in the early 1980s, well explains [Nagao 1992] that it surpasses the conventional one a method that depends on tedious work of linguists. As if they were lexicographers, they had to manually place semantic indicators at each noun, exactly indicate verb values and so on. That it is a difficult, expensive and protracted affair. On the other hand, a supply of examples suffices comprehensive to translate whole sentences based on it is simply impossible. Nagao and [Sato, 1990] guided researchers on the new path by describing a hybrid EBAT framework that makes it possible to integrate example phrases into the whole of the sentence syntactic structure. Notable in this proposed framework is the use of dependency trees instead of the hitherto usual constituent trees. Also in his second, full-sentence prototype Sato used dependency trees.

A decade later, [Yamamoto 2000] confirms the use of dependency syntax structures for alignment of phrases in PAT and implicitly in EBAT. That now helps solve a more general problem, which the US IBM Models 1-5 did not affect: the alignment of a source language sequence to only one target language word. The classic example of this is the English *'red herring'* and its German counterpart *'Finte'* , but there are plenty of such non-composite 41 translations. Connect (via only one connection line) whole word sequence from the source language to an entire word sequence in the target language, including IBM Models cannot, regardless of whether the word numbers in the two strings are equal or different. One would only think of idioms and idioms, precisely those phrases whose translation EBAT aims at.

In the research world, the hybrid translation machine, supporting both on PAT and on syntax, now gradually gains ground (PAT including EBAT, syntax including morphology). It is, however still fans who somewhat resist the return of syntax. [Koehn 2003] compared PAT-results in two variants of alignment: in one method all three-word sequences 42 were aligned, in the other only syntactic phrases 43 . The authors claimed a preference for syntactic structures made the translation worse, and they challenged the syntax-friendly peers.

[Lin 2004], who in the 1990s had explored powerful parsers with the help of syntax 44 , answered the challenge. While Koehn et al. based its syntactic variant on a constituent

40 *Word List by Semantic Principles* , NLRI (National Language Research Institute), Syuei Syuppan, Japan, 1964.
41 English: *'non-compositional'* .
42 in English: *'clump', 'word trigram'* or simply *'trigram'* .
43 in English: *'syntactic phrase'* or simply *'phrase'* .
44 see [Lin 1995].

Page 11

trees, Lin's learning translation machine extracts paths 45 from source language dependency trees of a word-aligned corpus, and translates those into fragments of a target language dependency tree. At the same time, not only the dependency relations, but also the linear sequence of words are encoded. Thus the corpus-a based learning process results in a set of transitional 46 rules with certain probabilities. Then, the translation of a new sentence develops here: parse the sentence to get its dependency tree; extract from that all paths, and find their translations; look for a combination of transition rules that handles the a source language tree completely and produces a target language dependency tree without conflict; if several such combinations are found, choose the one with the highest probability.
Lin's system went through a learning phase of 116,889 pairs of sentences (English-French, with 3.4 million words in total), from which 2,040,565 syntactic paths were extracted. The test phase contained 1775 sentences of 5-15 words length. Although the translation quality was still modest (BLEU-score: 0.26), promising is the agile transitional model whose syntax is capable of handling deviations such as the English-German pair *'there is'* - *'es gibt'* and the English-Spanish *'swim' across'* - *'cross swimming'* .

French trees are reviving, Americans are drying up

The new AT pathway also has the following characteristic: an increasing preference for dependency syntax. This is remarkable because for decades its great brother, the constituent syntax, ruled the AT-a world almost alone. Here, instead of the techniques I want to emphasize the almost cultural difference between the two.

Dependency syntax comes from the French Tesnière, in the mid-20th century, and acquired a certain adept among European linguists. But when in the United States the AT research developed, the Chomsky transformation-generative grammar greatly influenced the. local linguists. That model with its abstractions and constituent syntax became a real fad that penetrated also the circles of AT researchers in Europe and Japan. Judging by its publications, no dependency syntax existed at all. Here is the situation up to around the end of the 1980s.

It should not be forgotten that in the world of AT and computer linguistics in general, the English language has a prominent position. Most research, systems, corpora, parsers, software and so on concerns the English language. The most extensive knowledge and experience was accumulated about it. So, it is understandably and to some degree forgivable that such a prominent language is more or less reflected in the choice of methods and tools. A circumstance that has contributed to this is the fact that many English-speaking AT researchers, even the (modern) linguists among them, have only very limited knowledge about "foreign" languages. The simple use in English-language research publications of this epithet, to indicate other languages, reveals that.

While the constituent grammar is good enough for the English language, whose syntactic structure is based primarily on word order (constituent is actually a word sequence), it is less useful for languages with more morphology-based syntax. For language diversity, dependency is more appropriate. syntax, as this approaches contrastive syntax [Schubert 1986].

Tendency to dependency trees is undeniable. According to [Lopez 2002], the success of recent pars methods [Charniak 2000; Collins 1999; Ratnaparkhi 1999] is thanks to ideas basically inherent to dependency syntax. [Hwa 2002] confirms this and cleverly took advantage of a powerful parser for the English language that converts sentences into dependency trees. Such a parsley did not yet exist for Chinese. By word alignment between English and Chinese sentences of corpus [47] , Hwa (or more precisely: her learning machine) took word dependencies from the English side and projected them

45 *'paths'* in English terminology.
46 English *'transfer rules'* .
47 56,000 sentence pairs from Hong Kong News.

on the Chinese side, thus creating dependency trees there. By that experiment she showed that word dependencies are more suitable for interlanguage projection than the word constituents.

Constituent trees have not yet disappeared in AT, but on the current PAT road they have gradually disappeared loses its strength. [Knight 2004] admits that lineage distortion as in a phrase pair *'I had bought the car ' - ' Ich hatte das Auto gekauft '* cannot be handled without dependency syntax, and [Koehn 2002] reports on the need to limit sentence length to 6 words in a targeted experiment constituent-syntactic enrichment of PAT.

Finally, also as a bridge to semantics, dependency syntax works better than constituents. syntax. [Hwa 2002] states: *"semantic dependencies constitute a superset based on syntactic dependencies "* , and referring to [Baker 1997]:" in the field of lexical semantics, research on the relationships between syntactic elements on the one hand and higher-level concepts as an *actor, beneficiary, issue,* on the other hand, focused mainly on syntactic dependencies, not on constituents ".

DLT results are shown to be persistent

Looking back, to what extent does the then DLT project [48] relate to PAT ? That project that included ambitious research on AT in and out of Esperanto, in fact took place before the paradigm shift from around 1990, as well as its ten times larger competitor EUROTRA [49] .

It is all the more remarkable that DLT chief grammarian Klaus Schubert already wisely in the mid-1980s and boldly preceded the above-mentioned trend toward dependency syntax. In a period when that method was still generally ignored in the AT circles, he perceived it as most suitable for multilingual translation system, and published extensively on it [Schubert 1986, 1987].

As indicated above, the purpose of dependency syntax in AT is to facilitate the projection or transition of elements from a source language structure to that of the target language, hence a contrastive syntax or a *'metatax'* , as Schubert called - in Tesnière's honor - the process.
In addition, Schubert not only described and motivated the principles of metatestimation, but from 1986 to 1989 also co-organized the writing of concrete dependency syntaxes for 10 languages 50 . Also theirs results have been published [Maxwell 1989].

Schubert's chosen dependency syntax proved to be a solid foundation on which in the years 1987-1989 his colleague and chief semanticist at DLT, Victor Sadler, built an avant-garde method to enable a kind of EBAT (Sample-Based Machine Translation). The author himself titled this' *by analogy semantics* ' 51 and published his work in a book [Sadler 1989], which was often referenced in Japanese facts in the early 1990s. While the above-mentioned EBAT prototypes in Japan were used a thesaurus for calculating semantic 'distances' between words or phrases, **Sadler** 's method **only needs the text corpus itself, which in its entirety serves as an example base and thesaurus at the same time. This architecture put DLT on the threshold of the new (PAT) path. i-related, see also the overview in [Hutchins 1992].**

Semantic word distance, or *semantic proximity 52* , as Sadler called it, is the core of his invention. Do not confuse it with the linear distance between two words 53 , which equals the number of

48 Distributed Language Translation, a research project at the then Dutch software company BSO, during the years 1982-1990 (see: http://ourworld.compuserve.com/homepages/profcon/e_dlt.htm).
49 largest-scale AT research project ever, in which approx. three hundred university students from across Europe took part, and it was funded by the European Community (1978-1993).
50 languages and writers were: English (Bieke van der Korst, Dan Maxwell), Bengali (Probal Dasgupta), Danish (Ingrid Schubert), Esperanto (Klaus Schubert), Finnish (Kalevi Tarwainen), French (Luc Isaac, Dorine Tamis), German (Henning Lobin), Hungarian (Gábor Prószeky, Ilona Koutny, Balázs Wacha), Japanese (Shigeru Sato) and Polish (Marek Świdziński).
51 English *'analogical semantics'* .
52 English *semantic proximity* .
53 English *word 'co-occurrence'* .

Page 13

intermediate words plus 1, and which are used by the technology of search engines on the internet, sometimes even certain translation systems. However, to get a high-quality translation, the purpose of PAT, more subtle gear necessary. To fully understand the nature of *semantic proximity* , imagine that you urgently need it a complete picture of the meaning difference between two words, just as those two words are used in the a practice that is reflected in a large body of text. Neither a dictionary nor a separate thesaurus is available, so you asks for concordance, a list of all the contexts in which word number 1 appears 54 . If you have computer memory and speed, you immediately encapsulate that context. Then, you immediately turn on a concordance of word number 2, and in the next microseconds you sums up the differences between the two contexts and deduces from this the *semantic proximity* of the two words: a digit value to two decimal places 55 . Some examples:

government	board	0.89
government	federation	0.78
government	convention	0.64
government	communication	0.35
government	principle	0.27
government	garbo	0.11

Here, the novelty lies in a special definition of 'context': dependency-syntactic relations 56 with neighboring words, rather than purely linear proximities, even if the latter coincidentally coincide with the first. The formula by which Sadler in 1989 teaches the learning machine to calculate *semantics proximity is* therefore based on dependency relations, same as those 57 introduced by the above-mentioned researcher [Lin, 2004] more than a decade later. Enrichment of corpus by parsing, which the method of Sadler claims it is feasible because paraphrasing parsing - tracking single dependency relationships - enough.

Let us be aware that it is the corpus alone that causes the two-decimal semantic digits by the Sadler's dependency syntax measure. If by chance the corpus would be a novel in which people are constantly would say that they do not trust their husband, nor their government, that both the husband and the government are estranged money that for that reason they will rejoice in a change of husband and in a change of government, then the semantic proximity between the words 'government' and 'husband' could possibly reach a value of 0.90.
The corpus-based aspect, however, has a great advantage over the use of thesaurus, taxonomy or ontology 58 . Such encyclopedic structures not only require continuous updating (this corpus in fact also needs a corpus), but their maintenance implies: to specify in a hierarchy the place of each new addition. This is precisely - certainly in more abstract concepts - often risky and sometimes an undecidable matter. It is wise to take care of and equally expand the corpus as a single knowledge base is carefree, but at least doable.

The *semantic proximity* developed by Sadler, and linked to the dependency syntax provided by Schubert, is the leftover treasure of DLT. Its value is lasting and current, since in 1989 it sufficiently anticipated developments in the field, creating *"semantic dependencies based on syntactic dependencies "* (cf. [Hwa 2002]). Conveniently, the treasure is also accessible. It is extensively and in detail documented by publications 59 .

In conclusion, vetlude?

54 known in English as *'KWIC (KeyWord in Context)'* .
55 maximum 1.00 (in case of exact equality between the contexts).
56 subordinate relations, e.g. 'Verb - Object', 'Noun - Adjective', 'Preposition - Noun group'.
57 researcher He called these dependency relationships 'paths'.
58 copies. WordNet, EuroWordNet, and *"Word List by Semantic Principles"* (NLRI).
59 [Schubert 1986, 1987] are available at the UEA Book Service; at www.amazon.com available for purchase are [Schubert 1987], [Maxwell 1989] and [Sadler 1989].

Page 14

Looking ahead, at what amounts of money do we dare to bet that high-quality machine translation ready in 2020, or in 2030…?

Certainly the memory capacity of the future computers will not be problematic, nor the speed of those universal devices. Already they are enough for almost all machine translation tasks. Also the supply of text corpora (learning material for the learning translation machine) continues to grow and be updated. In fact the internet itself will increasingly function as a huge multilingual corpus, and a growing number of researchers use it in this way.

The new base of PAT, in a clever hybrid spider web with some sort of syntactic elements, looks healthy and promising. Compared to the rule-based rationalism of the traditional paradigm, which was too inclined to perfecting abstract language models, the current persististic and empiricist alpaca appears more daily for gradual and continuous improvement of translation machines.

Intuitively let's expect a statistical core to make a system more flexible, like a flavor for all those unforeseen counter-regular cases, including typos, unqualified proper names, quotations in other languages etc. The recent progress on the new path already shows that syntactic analysis of items, therefore to be combined with corpus-based statistics, is a more successful strategy than the eternal efforts to build a perfect parser that will impeccably find the only correct analysis of any whole sentence 60 . The persuasive operation certainly contains redundancy: several translations can result, even with negligible probability differences. This can reinforce the translation process.

On the other hand let us not forget that the learning machines in test phase so far translate only approx. 50 percent

of the sentences presented well. One way to progress is to expand the corpus. The more
the larger the database, the more reliable the statistics. Another resource is expansion and improvement of
the various processes (corpus preparation, alignment, parsing, transition, text structure analysis).

But the most critical factor on which the breakdown of a high-quality translation machine will depend will be organizational,
not technological! The researchers, scattered through universities, naturally inclined to create ever new variants,
rarely commits itself to a common further construction of one system. The business finds high quality
a general translation system is not attractive enough, and an international government like the one in Brussels is not
dare (again) to risk a large expense for it. It takes pressure and outstanding organization to do that
competent fans by joint forces carry out difficult multi-year cooperation. How
expressed a long-time AT researcher [Carbonell 1992]: *"in Atomic Translation, persistence
it matters "* .

A rebirth of a (P) AT project in Esperanto, as a successor to DLT, isn't that worth betting on?
Maybe an international network or a miraculous local grouping of language-conscious computer scientists…
experts for whom committed cooperation will temporarily make up for the lack of
employment contract in Hyderabad power station?

Bibliography

[Baker 1997] Mark C. Baker. *Thematic Roles and Syntactic Structure* . Kluwer. p. 73–137.

[Brown 1988] Peter F. Brown et al .: *A statistical approach to language translation.* Proceedings International Conference on Computational Linguistics (COLING-88). Budapest. p. 71-76.

60 Even for the English, of which the gear is most advanced, four decades of AT history a decent parser was not available.

Page 15

[Brown 1990] Peter F. Brown et al .: *A statistical approach to language translation.* Computational Linguistics, June 1990, vol. 16, No. 2, p. 79-85.

[Brown 1993] Peter F. Brown et al .: *The mathematics of Statistical Machine Translation: Parameter Estimation.* Computational Linguistics, June 1993, vol. 19, No. 2, p. 263-311.

[Callison-Burch 2004] Chris Callison-Burch, Colin Bannard, Josh Schroeder: *Improved Statistical Translation Through Editing.* School of Informatics, University of Edinburgh; Linear B Ltd., Edinburgh Technology Transfer Center.

[Carbonell 1992] Jaime G. Carbonell, Teruko Mitamura, Eric H. Nyberg, 3 rd : *The KANT Perspective: A Critique of Pure Transfer (and Pure Interlingua, Pure Statistics,...).* Fourth International Conference on Theoretical and Methodological Issues in Machine Translation, Montreal.

[Charniak 2000] Eugene Charniak: *A maximum-entropy-inspired parser.* Proceedings 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL), Seattle.

[Collins 1999] Michael Collins: *Head-Driven Statistical Models for Natural Language Parsing* . Ph.D. thesis, University of Pennsylvania.

[Ding 2003] Yuan Ding, Daniel Gildea, Martha Palmer: *An Algorithm for Word-Level Alignment*

of Parallel Dependency Trees . Proceedings MT Summit IX, New Orleans.

[Hacken 2001] Pius ten Hacken: *Revolution in Computational Linguistics.* Language and Computers, December 2001, vol. 37, No. 1, p. 60-72 (13).

[Hutchins 1992] W. John Hutchins, Harold L. Somers: *An Introduction to Machine Translation.* Academic Press.

[Hutchins 2003] John Hutchins: *Has machine translation improved? Some historical comparisons.* Proceedings MT Summit IX, New Orleans.

[Hwa 2002] Rebecca Hwa, Philip Resnik, AmyWeinberg: *Breaking the Resource Bottleneck for Multilingual Parsing.* Institute for Advanced Computer Studies and Department of Linguistics, University of Maryland.

[Knight 2004] Kevin Knight, Philipp Koehn: *What's New in Statistical Machine Translation* . Tutorial at Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Boston.

[Koehn 2002] Philipp Koehn, Kevin Knight: *ChunkMT: Statistical Machine Translation with Richer Linguistic Knowledge.* By: http://people.csail.mit.edu/people/koehn.

[Koehn 2003] Philipp Koehn, Franz Josef Och, Daniel Marcu: *Statistical Phrase-Based Translation.* Proceedings (Main Papers) Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Edmonton. p. 48-54.

[Lin 1995] Dekang Lin: *A dependency-based method for evaluating broad-coverage parsers.* Proceedings International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal. p. 1420-1425.

[Lin 2004] Dekang Lin: *A Path-based Transfer Model for Machine Translation.* Proceedings International Conference on Computational Linguistics (COLING-2004), Geneva.

[Lopez 2002] Adam Lopez, Michael Nossal, Rebecca Hwa, Philip Resnik: *Word-level Alignment for Multilingual Resource Acquisition* . Language and Media Processing Laboratory

Page 16

(LAMP) Technical Report 085, Institute for Advanced Computer Studies, University of Maryland (UMIACS).

[Maxwell 1989] Dan Maxwell, Klaus Schubert (eds.): *Metataxis in Practice - Dependency syntax for multilingual machine translation.* Foris Publications.

[Nagao 1992] Makoto Nagao: *Some Rationales and Methodologies for Example-based Approach.* Proceedings, International Workshop on Fundamental Research for the Future Generation of Natural Language Processing (FGNLP). Sofia Ananiadou (ed.), Manchester.

[Ratnaparkhi 1999] Adwait Ratnaparkhi: *Learning to parse natural language with maximum entropy models.* Machine Learning, 34 (1-3) p. 151–175.

[Sadler 1989] Victor Sadler: *Working with Analogical Semantics: Disambiguation Techniques in DLT.* Foris Publications.

[Sato 1990] Satoshi Sato, Makoto Nagao: *Towards Memory-based Translation.* Proceedings, International Conference on Computational Linguistics (COLING-90), Helsinki.

[Sato 1991] Satoshi Sato: *Example-Based Machine Translation.* Ph.D. thesis, September 1991, Kyoto University.

[Schubert 1986] Klaus Schubert: *Syntactic Tree Structures in DLT.* BSO / Research, Utrecht.

[Schubert 1987] Klaus Schubert: *Metataxis - Contrastive dependency syntax for machine translation.* Foris Publications.

[Yamamoto 2000] Kaoru Yamamoto, Yuki Matsumoto: *Acquisition of Phrase-level Bilingual Correspondence using Dependency Structure.* Proceedings, International Conference on Computational Linguistics (COLING-2000), Saarbrücken.