Statistical and Adaptive Natural Language Processing

T-61.5020 (5 cr) L, Spring 2008

Sixth lecture

Machine Translation and Text Alignment

Lecturer: **Jaakko Väyrynen**

Slides: Krista Lagus, Mathias Creutz, Timo Honkela, Jaakko Väyrynen

6. Machine Translation

Lecture based on:

- Chapter 13-13.1 in Manning & Schütze.
- ‘Machine Translation: A Brief History’ by W. John Hutchins
- Dekai Wu. ’MT model space: statistical vs. compositional vs. example-based machine translation’. Machine Translation (2005) **19**:213–227. Springer

6.1 Machine Translation

- **Machine Translation** (MT): Computerized systems responsible for the production of translation with or without human assistance. It excludes tools witch only support access to dictionaries, terminology databanks, etc.
- CAT = Computer Aided Translation, HAMT = Human Aided Machine Translation, MAHT = Machine Aided Human Translation, L10N = Localisation
- Automatic machine translation is one of the oldest goals in language technology. It is, however, a very difficult problem.
- Current MT software produces mainly raw translations, that may speed up the work of a human translator, but are not necessarily good enough for human readers (post-editing is the norm).
- For very limited applications (such as weather forecasts) reasonable outcomes may be attained; English-French translation in Canada and Finnish-Swedish translation in Finland.

6.2 When does MT work best?

- Restricted input:
  Sublanguage (limited vocabulary and grammar), particular subject field (e.g. biochemistry), document type (e.g. patents)

- Input text in controlled language:
  Limited vocabulary, avoid ambiguity (homonymy, polysemy) and complexity

- Pre-edited text (markup):
  Indicators for prefixes, phrases, grammatical categories, etc.

- Interactive systems:
  System may refer with, e.g., ambiguities to operators

6.3 When does MT work worst?

- Literature, poems, cartoons, sayings, changes in language
- World knowledge: What is the color blue? What is the Grand Canyon?
- With different styles: conversational language, slang, SMS messages
- Stylistic choices, different levels of meaning in text
- 'Gadsby' by E. V. Wright, a 50,000 word story without the vowel 'e'
- 'the quick brown fox jumps over the lazy dog'
- 'hae lakkaa satamasta kun lakkaa satamasta'
- Mieleni minun tekevi,
  aivoni ajattelevi
  lähteäni laulamahan,
  saa'ani sanelemahan
- 'The rain in Spain stays mainly in the plain',
  'Vie fiestaan hienon miekkamiehen tie'

Different levels of translation

- The simplest approach, **word-to-word translation**, replaces words in the source language with words from the target language. The word order of the result is typically wrong.

- **Transfer systems** (syntactic and semantic) build a structured intermediate representation from a word sequence in the source language; transfer it to an intermediate representation in the target language (using some kind of rules); and generate a word sequence in the target language (analysis, transfer, generation).

- **Syntactic transfer** builds a syntactic structure representation from a word in sequence in the source language. The approach requires a working syntactic disambiguation.

  This solves word order problems, but often the result is not semantically correct. E.g., 'Ich esse gern' would translate to 'I eat readily' (or 'willingly', 'gladly'). However, there is no corresponding verb-adverb pair in English, the correct translation would be 'I like to eat'.

- **Semantic transfer** makes a more complete representation, semantic analysis, that should produce a translation which would also be semantically correct.

  However, semantically 'correct' translation might be clumsy in the target language, even if is basically understandable. E.g., the exact translation for the Spanish sentence 'La botella entró a la cueva flotando' would be 'the bottle entered the cave floating', but it would be more natural to say 'the bottle floated into the cave'.

  Often clumsy and unnatural translations slow down understanding, even if understanding would be possible in principal. An unnatural translation might also be more easily misinterpreted due to a possible ambiguity.

  By the way, this is a classical example of how the *manner* and *direction* of motion are expressed differently in different languages: In English the verb indicates the manner, whereas the satellite (added words) indicates the direction: *crawl out, float off, jump down, walk over to, run after*. In Spanish, the opposite is true: *salir flotando* (exit floating

= float out), *acercarse corriendo* (approach running = run towards), *alcanzar andando* (reach walking = reach by foot = walk all the way to).

Curiously enough, there also exist languages, where the verb describes *the shape of the one who moves or is moved*: something like *to slither* (a snake?).

- **Interlingua** – a common artificial (language independent) intermediate language or knowledge representation. Translation from the source language to the interlingua and from the interlingua to any target language.

  For n languages, instead of n^2 translators, only $2n$ translators are needed. In addition, they can be implemented partly using general language processing methods. However, defining a sufficient intermediate language is by itself a hard problem, that so far hasn't been solved in general.

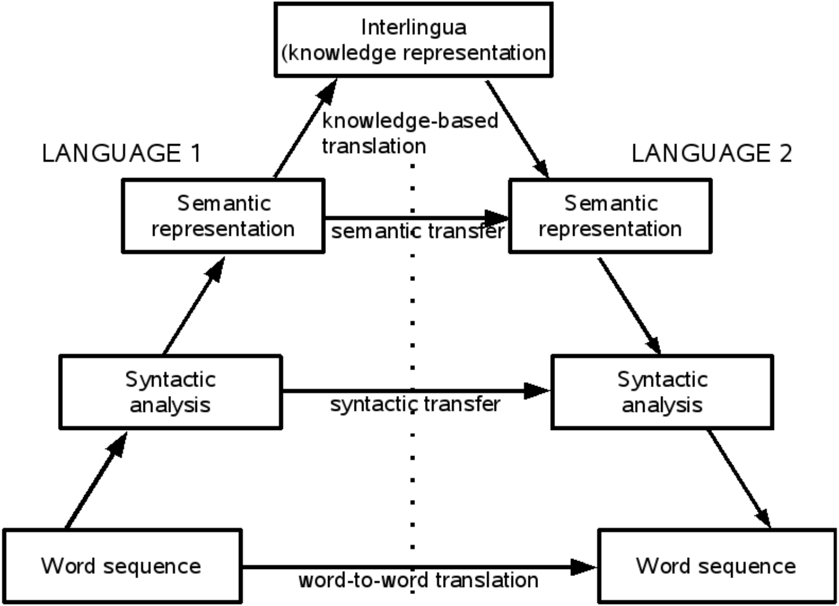
  There are practical problems in designing an efficient and comprehensive knowledge representation, as well as in ambiguity resolution.

Examples of interlingua:
Natural languages: English, South-American language Aymara;
Constructed languages: Esperanto, lojban;
Logical/Artificial languages: 1st order predicate logic.

- The image in the next page shows the alternative approaches to machine translation.

- Statistical language processing methods can be used as components for the system for any arrow (e.g., parsing, disambiguation, etc.).

- Machine translation systems may also be combinations of symbolic and statistical components.



Interlingua
(knowledge representation

knowledge-based
translation

LANGUAGE 1

LANGUAGE 2

Semantic
representation

semantic transfer

Semantic
representation

Syntactic
analysis

syntactic transfer

Syntactic
analysis

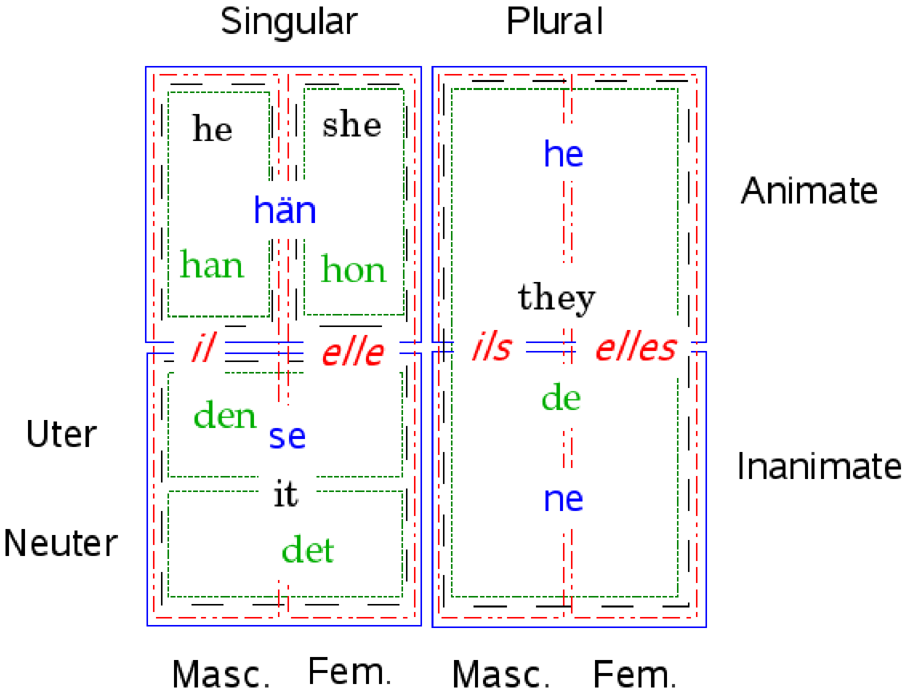
Word sequence

word-to-word translation

Word sequence

Semantic differences across languages

- Distinctions are made in different ways in different languages. That is, the “semantic space” is divided differently. For instance, if we are to translate the English color word *blue* into Russian, we have to know whether it is dark blue (*sinij*) or light blue (*goluboj*). (There is no word for “just blue” in Russian, but if we have to choose without any additional knowledge, we would probably take *sinij*.)
- Also compare some third-person pronouns in four languages. Distinctions are made based on number (singular, plural), animacy (animate, inanimate), and gender (masculine, feminine, uter, neuter):
  - Finnish: *hän, se, he, ne* (no gender distinction)
  - Swedish: *han, hon, den, det, de* (animacy and gender distinction only in singular)
  - English: *he, she, it, they* (animacy and gender only in singular)
  - French: *il, elle, ils, elles* (no animacy distinction)



Singular

Plural

he

she

hän

han

hon

he

they

Animate

il

elle

ils

elles

Uter

den

se

de

it

ne

Inanimate

Neuter

det

Masc.

Fem.

Masc.

Fem.

- Cultural and linguistic differences: **Sapir-Whorf hypothesis**: language constrains thought – the language you speak may affect the way you think.

  Some things are left ambiguous in one language, whereas they have to be fixed in another. For instance, in Russian you have to decide between dark and light blue, and in many languages you have to decide between "he" and "she". (Does this mean that Finnish speakers are more likely to think of persons gender-neutrally? Does it mean that French speakers associate masculine and feminine properties with inanimate objects? Does a plant (*une plante*) have female features, whereas a tree (*un arbre*) has masculine ones?)

- To the extent that the Sapir-Whorf hypothesis is true, there can be no perfect translation, since speakers of the source and target languages necessarily have different conceptual systems.

- Couldn't interlingua be the solution: an artifical language with very fine-grained semantic categories and distinctions?

  - Problem 1: How construct such an intrinsically complicated system?

  - Problem 2: If two languages are rather closely related, they may share a number of constructs and ambiguities. If we were to translate the Swedish sentence “De kom för sent.” into English, we could actually do almost with a word-to-word translation: “They came too late.”. It does not matter whether “they” are animate or inanimate (“he ihmiset” or “ne taksit”) or whether they came walking, swimming, or driving. It does not seem optimal to require a parser to perform deeper analysis and more disambiguation than necessary.

- Before moving on to statistical methods, we shall take a brief look at two examples of **rule-based direct translation**, which does not make use of complex structures and representations (no interlingua). The input is treated as a string of words (or morphemes), and various operations are performed directly on it.

Japanese-to-English Translation

Stage	Action
1.	morphological analysis
2.	lexical transfer of content words
3.	various work relating to prepositions
4.	SVO rearrangements
5.	miscellany
6.	morphological generation

Input:	watashihatsukuenouenopenwojonniageta.
After stage 1:	watashi ha tsukue no ue no pen wo jon ni ageru PAST.
After stage 2:	*I* ha *desk* no ue no *pen* wo *John* ni *give* PAST.
After stage 3:	I ha pen on desk wo John to give PAST.
After stage 4:	I give PAST pen on desk John to.
After stage 5:	I give PAST the pen on the desk to John.
After stage 6:	I gave the pen on the desk to John.

Direct translation of ‘much’ and ‘many’ into Russian

```
if preceding word is how /* how much, how many */
    return skol’ko
else if preceding word is as /* as much, as many */
    return stol’ko zhe
else if word is much
    if preceding word is very /* very much */
        return nil (not translated)
    else if following word is a noun /* much people, food */
        return mnogo
else /* word is many */
    if preceding word is a preposition and following word is a noun
        return mnogii
else return mnogo
```

Adapted from Hutchins’ (1986) discussion of Panov (1960).

Imagine that you’d have to debug such a system with all its word-specific rules...

6.4 Short History of Machine Translation

- 1950s and 1960s: Pioneers and early systems.
- mid 1960s: ALPAC report.
- 1970s: Revival.
- 1980s: Commercial and 2nd generation systems, research.
- 1990s: New developments in research, growing use of systems.

Precursors and Pioneers: 1933–1956

- 1933: First patent proposals.
- 1949: Suggestions of applying cryptography techniques, Shannon's information theory, statistical methods and 'logic underlying the language' by Warren Weaver.
- 1952: First MT conference.
- 1954: First public demonstration of an MT system.
  - 49 Russian sentences into English.
  - 250 word vocabulary.
  - 6 grammar rules.

High expectations and disillusions: 1956–1966

- Focus on two different approaches: empirical systems with statistical methods and theoretical linguistic solutions.

- Developments:
  improvement of basic computer facilities (memory, speed, . . .),
  programming tools for language processing,
  parser based on dependency grammar by David Hays.

- Systems:
  focus on Russian↔English,
  word-to-word direct system with bi-lingual lexicons,
  morphological, syntagmatic and syntactic analysis,
  syntactic transfer,
  interlingua, semantic networks of a thesaurus.

- Finally hit the semantic barrier.

ALPAC report

- Complexity of the linguistic problems became more apparent.
- 1966: Automatic Language Processing Advisory Committee (ALPAC) concluded that MT was slower, less accurate and twice as expensive as human translation and that there was no need to continue research on MT.
- Virtual end to MT research in the USA for a decade.
- MT activity switched to Canada and the European Communities, which had a growing need for translations of administrative and legal texts.

The quiet decade: 1966–1976

- Main approaches: syntactic transfer, interlingua.
- Developments by the TAUM project at Montreal:
  Q-system formalism, a foundation to Prolog.
- Systems:
  Météo system for meteorological reports by TAUM,
  the 'pivot language' of CETA,
  TITUS for abstracts (multilingual, in France),
  CULT (Chinese-English, in Hong Kong),
  Systran (Russian-English, in US Air Force).
- Interlingua systems provided disappointing results due to:
  rigidity of the levels of analysis,
  inefficiency of parsers,
  loss of information about the surface forms of SL input.

Commercial and 2nd generation systems: 1976–1989

- Research in:
  advanced transfer systems,
  new kind of interlingua systems,
  investigation of techniques from artifical intelligence (AI).
- Developments in:
  knowledge-based systems,
  modularity of 2nd generation systems,
  Eurotra project (much research, no prototype).
- Systems:
  Systran (English-Russian, in EC/EU),
  Logos (German-English),
  PENSEE, MELTRAN, ATLAS, . . . (English↔Japanese),
  METAL (German-English),
  Weidner and ALPS (for microcomputers)
- However, commercial systems are (originally) 1G systems.

Corpus-based MT research: 1989–

- Revival of statistical approach by large parallel corpora and accomplishments of speech recognition research.
- 1988: IBM published a paper on a purely statistical method for MT.
- Candide project at IBM: French-English.
- At the same time the birth of example-based MT in Japan.

Rule-based MT research: 1990–

- Research in transfer and interlingua systems.
- Systems and projects:
  Linguistic-based transfer: Eurolang, LMT
  Knowledge-based interlingua: CATALYST
  Linguistic-based interlingua: UNITRAN

New areas of research: 1990–

- Generation of good quality texts.
- Dialogue-based MT.
- Spoken language translation.
- Transportable translation tools.

Operational systems: 1990–

- Mainframe systems for commercial agencies.
- Translation tools and workstations.
- Systems for personal computers, online translation.

Summary

- Several approaches, none has proved best.
- Hybrid systems blur differences between paradigms.
- Most commercial systems are originally 1G systems: Systran, used by EU, Google and Microsoft.
- Demand for different kinds of translation systems

6.5 Dimensions of Machine Translation Paradigms

- **Statistical MT**, making nontrivial use of mathematical statistics and probability, vs.

- **Logical MT**, making nontrivial use of mathematical logic.

- **Example-based MT**, making nontrivial use of a large library of examples during translation runtime, vs.

- **Schema-based MT**, making nontrival use of abstract schemata during translation runtime.

- **Compositional MT**, making nontrivial use of compositional transfer/transduction rules, vs.

- **Lexical MT**, making nontrival use of lexical transfer/transduction rules.

- Another viewpoint at the division of translation systems.

- Translation dictionaries (not MT by definition).

- **Rule-based MT** (RBMT).

- Corpus-based MT (CBMT) can be divided into:
  Example-based MT (EBMT): {Exact, fuzzy} match, translation by analogy, and
  Statistical MT (SMT): Probabilistic match, language and translations models.

- The success of corpus-based machine translation depends heavily on the quality of the **text alignment** that is produced.

6.6 Example systems

Rule-based Danish-English (by Eckhard Bick)

- Danish Constraint Grammar with rules for morphological and PoS disambiguation, mapping and disambiguation of syntactic functions (6,000 rules).
- Dependency rules establishing syntactic-semantic links between words or multi-word expressions (220 rules).
- Lexical transfer rules, selecting translation equivalents according to grammatical category, dependency and other structural context (17,000 rules).
- Generation rules for inflection, verb chains, compositions etc. (700 rules).
- Syntactic transformation (movement) rules to establish English word order, handle subclauses, negation, questions etc. (75 rules).
- `http://gramtrans.com/`.

Statistical phrase-based Finnish-English

- Trained on a parallel corpus (8.6M sentence pairs).
- Translation and reordering scores (14M phrase pairs).
- 4-gram language model for English (5.5M n-grams).
- `http://cog.hut.fi/smtdemo`.

6.7 Text Alignment

- Text alignment means aligning **parallel texts** (or a bi-text) in different languages, so that the corresponding text fragments are matched.
- Parallel texts contain the same document in different languages.
- Most often used parallel texts are administrative texts from countries or confederations with several official languages (e.g., Europarl, Canadian and Hong Kong Hansards, KOTUS Finnish-Swedish parallel corpus).

  In addition to public availability, administrative parallel texts are usually translated consistently and as exactly as possible. Such high quality material is important for both the development of statistical methods as well as for the evaluation of the methods.

- Also newspapers and magazines are sometimes used, and also religious texts would be easily available. The results, however, are typically inferior, presumably from less exact and consistent translations and less static styles of text (e.g., current news topics change rapidly)

- There are typically two stages in text alignment

    1. Sentence and paragraph alignment: coarse text alignment, where matching paragraphs, sentences and sentence pairs are approximately aligned.
    2. Word alignment and bi-lingual dictionary extraction, where based on coarsely aligned material, the equivalents for source language words (and phrases) and found in the target language.

Sentence and paragraph alignment

Typically sentence alignment is a necessary first step in producing a multilingual corpus.

In addition to machine translation and bilingual dictionary induction, alignment can also benefit other applications, such as

- Word sense disambiguation (WSD): Word senses can be grouped according to its the equivalents. For instance, the Finnish word *kuusi* can be translated as *six* or *spruce* (or *your moon*).
- Multilingual information retrieval: Information source and query can have different languages.
- Tool for translator: As one document changes, automatically point out the location that must be updated in the other document in another language, and perhaps also propose the update.

Beading

A **bead** (jyvä) is a sentence or a sequence of a few sentences and the corresponding sequence of sentences (in the aligned text). Either of the beads can also be empty. Each sentence is part of exactly one bead.

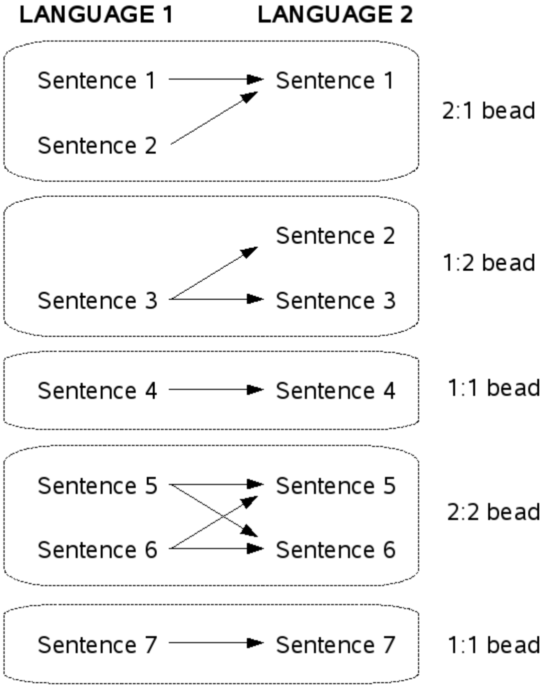
Beading (jyvitys) is a mapping, in which a bi-text is separated into fragments, and that tells which fragment in the source language corresponds to which fragment in the target language.

Sentence alignment is not a trivial problem, as for a sentence in the source language there isn't nowhere near always exactly one corresponding sentence in the target language (1:1 bead).

1:2 and 2:2 beads (also 1:3 and 3:1): Sentences are segmented differently. A human translator applies different orderings to produce a natural end result.

In a 2:2 bead, parts of two consecutive sentences in the source language are presented in two consecutive sentences in the target language (sufficient overlap).

When is the overlap sufficient? Typically crossover of just a couple of words is not enough, but overlap of a full sentence is expected.


LANGUAGE 1
LANGUAGE 2
Sentence 1
Sentence 1
Sentence 2
2:1 bead
Sentence 2
Sentence 3
Sentence 3
1:2 bead
Sentence 4
Sentence 4
1:1 bead
Sentence 5
Sentence 5
Sentence 6
Sentence 6
2:2 bead
Sentence 7
Sentence 7
1:1 bead

Example of a 2:2 Alignment

The sentence divisions in the English and French texts are different:

- *English:* According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products.
  Cola drink manufacturers in particular achieved above average growth rates.

- *French:* Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes.
  En effet notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment.

- *French-to-English literal translation:* With regard to the mineral waters and the lemonades, they encounter still more users.
  Indeed our survey makes stand out the sales clearly superior to those in 1987 for cola-based drinks especially.

Deletions and insertions a.k.a. 1:0 and 0:1 beads:

Some facts can be explicitly stated in one language but can be left out in the other language because they are expected to be implicitly interpreted (possible causes: word order, sense distribution, culture, expected knowledge of the target reader, etc.).

From different studies it can be estimated that roughly 90% of beads are of type 1:1 (however, the rate probably depends on the language pair and style).

It is also common that translators reorder sentences in different ways. The models mentioned here are not able to represent this, but interpret such cases, for instance, as insertions and deletions.

Statistical methods for text alignemnt

Some of the statistical methods are based solely on the examination of the lengths of text fragments. Whereas other take into the account the used lexicon (strings).

- Methods based on the length of text fragments
- Methods based on identical strings
- Lexical methods

From now on: let text F in language 1 be a sequence of sentences $F = (f_1, \ldots, f_I)$ and similarly for text E in language 2, $E = (e_1, \ldots, e_J)$ (F = Foreign, E = English)

Methods based on the length of text fragments

Several of the earlier text alignments methods are of this type.

Find the alignment A which has the largest probability:

$$\arg\max_A P(A|F,E) = \arg\max_A P(A,F,E) \tag{1}$$

(the most likely alignment can be found with, e.g., dynamic programming).

Several methods divide the aligned text into a sequence of beads $(B_1,\ldots,B_K)$ and approximate the probability of the whole aligned text by assuming that the probability of a bead does not depend on other sentences or beads, but only on the sentences in the bead in question:

$$P(A,F,E) = \prod_{k=1}^{K} P(B_k) \tag{2}$$

Bead probability calculation

Gale & Church, 1991, 1993:

The probability of a bead depends on the length of the bead's sentences measured in characters. The method is thus based on the assumption that a long fragment in one language is likely to correspond to a long fragment in the other language.

Let us assume that the bi-text is already aligned at paragraph level (for computational efficiency)

Only beads $\{1:1,\ 1:0,\ 0:1,\ 2:1,\ 1:2,\ 2:2\}$ are allowed.

Let $D(i,j)$ be the found smallest cost alignment between sentences $f_1,\ldots,f_i$ and $e_1,\ldots,e_j$.

We calculate $D(i,j)$ recursively. For the basic case, define $D(0,0) = 0$.

Recursion:

$$D(i,j) = \min \begin{array}{ll} D(i,j-1) & +cost(0:1 \text{ bead } 0, e_j) \\ D(i-1,j) & +cost(1:0 \text{ bead } f_i, 0) \\ D(i-1,j-1) & +cost(1:1 \text{ bead } f_i, e_j) \\ D(i-1,j-2) & +cost(1:2 \text{ bead } f_i, e_{j-1}, e_j) \\ D(i-2,j-1) & +cost(2:1 \text{ bead } f_{i-1}, f_i, e_j) \\ D(i-2,j-2) & +cost(2:2 \text{ bead } f_{i-1}, f_i, e_{j-1}, e_j) \end{array}$$

The cost of each bead type is calculated as follows:

Assume underlying model: one character in language L_1 generates a random number of characters in L_2. Assume that the number of generated characters is distributed as a Gaussian. The mean μ and variance σ^2 of the distribution are estimated from a large parallel corpus (for German/English pair it was estimated that $\mu = 1.1$, for French/English pair it was estimed that $\mu = 1.06$)

As a cost for the model one can use the the negative log-likelihood of the lengths of the text fragments.

$$cost(l_1, l_2) = -\log P(\alpha \text{ bead } |\delta(l_1, l_2, \mu, \sigma^2) \tag{3}$$

in which α is one of the allowed bead types and δ measures the difference between character lengths to the mean and variance estimates from the whole text: $\delta(l_1, l_2, \mu, \sigma^2) = (l_2 - l_1\mu)/\sqrt{l_1\sigma^2}$.

The requires probabilities are estimed by applying the Bayes formula

$$P(\alpha|\delta) = P(\alpha)P(\delta|\alpha) \tag{4}$$

In which case the high *a priori* probability $(P(\alpha = 1 : 1) = 90\,\%)$ favors choosing that bead.

Recursive cost calculation algorithm is slow if the texts are long. With single paragraphs, however, it is reasonably fast.

The methods words quite well with related languages: reported 4% errors. With the additional aim of separately detecting suspicious alignments and alignment of only the best 80% an error rate of 0.7% was reached.

The methods works best with 1:1-beads (2%), but with more complicated aligments error rates are high.

Brown et al 1991:

A variant of the previous method counts sentence lengths in words rather than characters. It has been argumented that this is a worse approach as there is more variance in word counts than in character counts.

Church, 1993: Method based on identical strings

The previous methods are not well suited for noisy text (e.g., created with optical character recognition), in which there might be gargabe in between sentences or fully missing paragraphs. Also paragraph and sentence borders are harder to detect because of, e.g., missing punctuation marks or nonsense text.

The observation at the foundation of this method:

In texts which are written in roughly the same alphabet (e.g., the Roman alphabet), there are identical strings suchs as proper names and numbers with the same meaning.

With related languages or languages that have had close interaction with each other, there can also occur other common strings as a consequence of shared origins (e.g., 'superior' in English and 'supérieur' in French) or loanwords.

Count identical character n-grams (e.g., with $n = 4$). Find the n-gram alignment that contains the maximum amount of identical n-gram pairs. In addition, the n-grams can be weighted with their frequency.

The method does not produce actual beading of sentences.

Can fail completely if there aren't enough shared strings.

Lexical methods

The aim is to produce a real sentence level beading.

It seems clear that knowledge about probable word pair translations would help the alignment considerably.

The main idea for purely statistical methods:
Alternate between probable partial alignment at the word level and the most probable sentence level alignment.

In addition it is assumed that equivalent sentence sequences are not too far from each other (e.g., no crossovers or not too far).

Only a few iterations are typically required (because of the constraint above).