Machine Translation

Berlin Chen 2004

References:

1. Natural Language Understanding, Chapter 13
2. W. A. Gale and K. W. Church, A Program for Aligning Sentences in Bilingual Corpora, Computational Linguistics 1993
3. Pattern Recognition in Speech and Language Processing, Chapter 11

Machine Translation (MT)

- Definition
  - Automatic translation of text or speech from one language to another
- Goal
  - Produce close to error-free output that reads fluently in the target language
  - Far from it ?
- Current Status
  - Existing systems are used in restricted domains
    - E.g. weather reports
  - A mix of probabilistic and non-probabilistic components

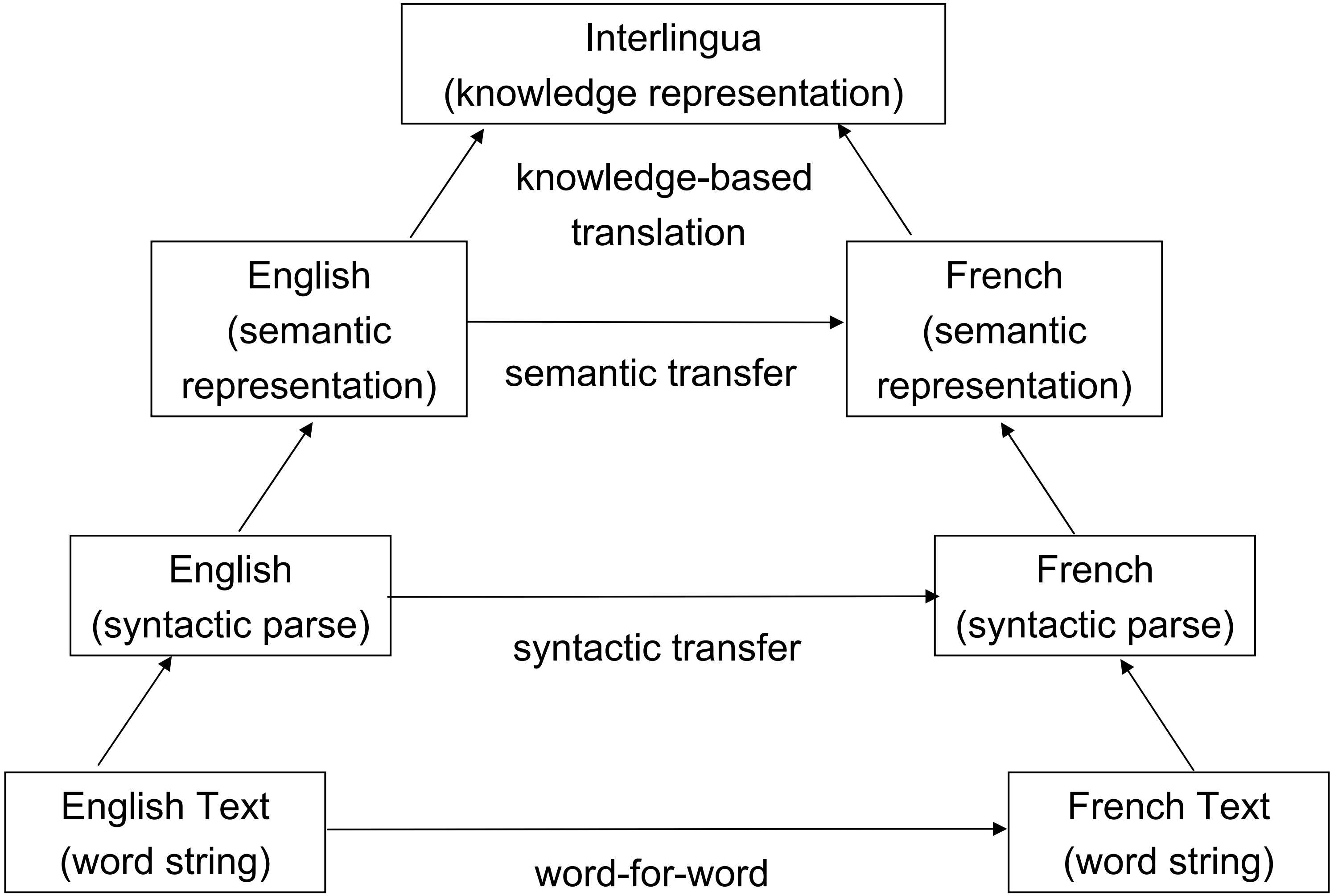
Issues

- Build high-quality semantic-based MT systems in circumscribed domains

- Abandon automatic MT, build software to assist human translators instead
  - Post-edit the output of a buggy translation

- Develop automatic knowledge acquisition techniques for improving general-purpose MT
  - Supervised or unsupervised learning

Different Strategies for MT



Interlingua
(knowledge representation)

knowledge-based
translation

English
(semantic
representation)

semantic transfer

French
(semantic
representation)

English
(syntactic parse)

syntactic transfer

French
(syntactic parse)

English Text
(word string)

word-for-word

French Text
(word string)

Word for Word MT

1950

- Translate words one-by-one from one language to another
  - Problems
    1. No one-to-one correspondence between words in different languages (lexical ambiguity)
       - Need to look at the context larger than individual word (→ phrase or clause)
    2. Languages have different word orders

English	**French**
suit	lawsuit, set of garments

meanings

Syntactic Transfer MT

- Parse the source text, then transfer the parse tree of the source text into a syntactic tree in the target language, and then generate the translation from this syntactic tree
  - Solve the problems of word ordering
  - Problems
    - Syntactic ambiguity
    - The target syntax will likely mirror that of the source text

N V Adv

German: Ich esse gern (*I like to eat*)

English: I eat readily/gladly

Semantic Transfer MT

- Represent the meaning of the source sentence and then generate the translation from the meaning
  - Fix cases of syntactic mismatch
  - Problems
    - Still be unnatural to the point of being unintelligible
    - Difficult to build the translation system for all pairs of languages

Spanish: La botella entró a la cueva flotando
(The bottle floated into the cave)

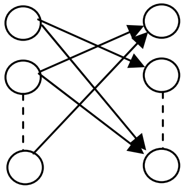
English: The bottle entered the cave floating

(In Spanish, the direction is expressed using the verb and the manner is expressed with a separate phrase)

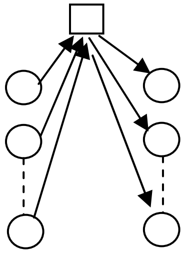
Knowledge-Based MT

- The translation is performed by way of a knowledge representation formulism called "interlingua"
  - Independence of the way particular languages express meaning

- Problems
  - Difficult to design an efficient and comprehensive knowledge representation formulism
  - Large amount of ambiguity needed to be solved to translate from a natural language to a knowledge representation language



n(n-1)



2n

Text Alignment: Definition

- Definition
  - Align paragraphs, sentences or words in one language to paragraphs, sentences or words in another languages
    - Thus can learn which words tend to be translated by which other words in another language

      bilingual dictionaries, MT , parallel grammars …
  - Is not part of MT process per se
    - But the obligatory first step for making use of multilingual text corpora

- Applications
  - Bilingual lexicography
  - Machine translation
  - Multilingual information retrieval
  - …

Text Alignment: Sources and Granularities

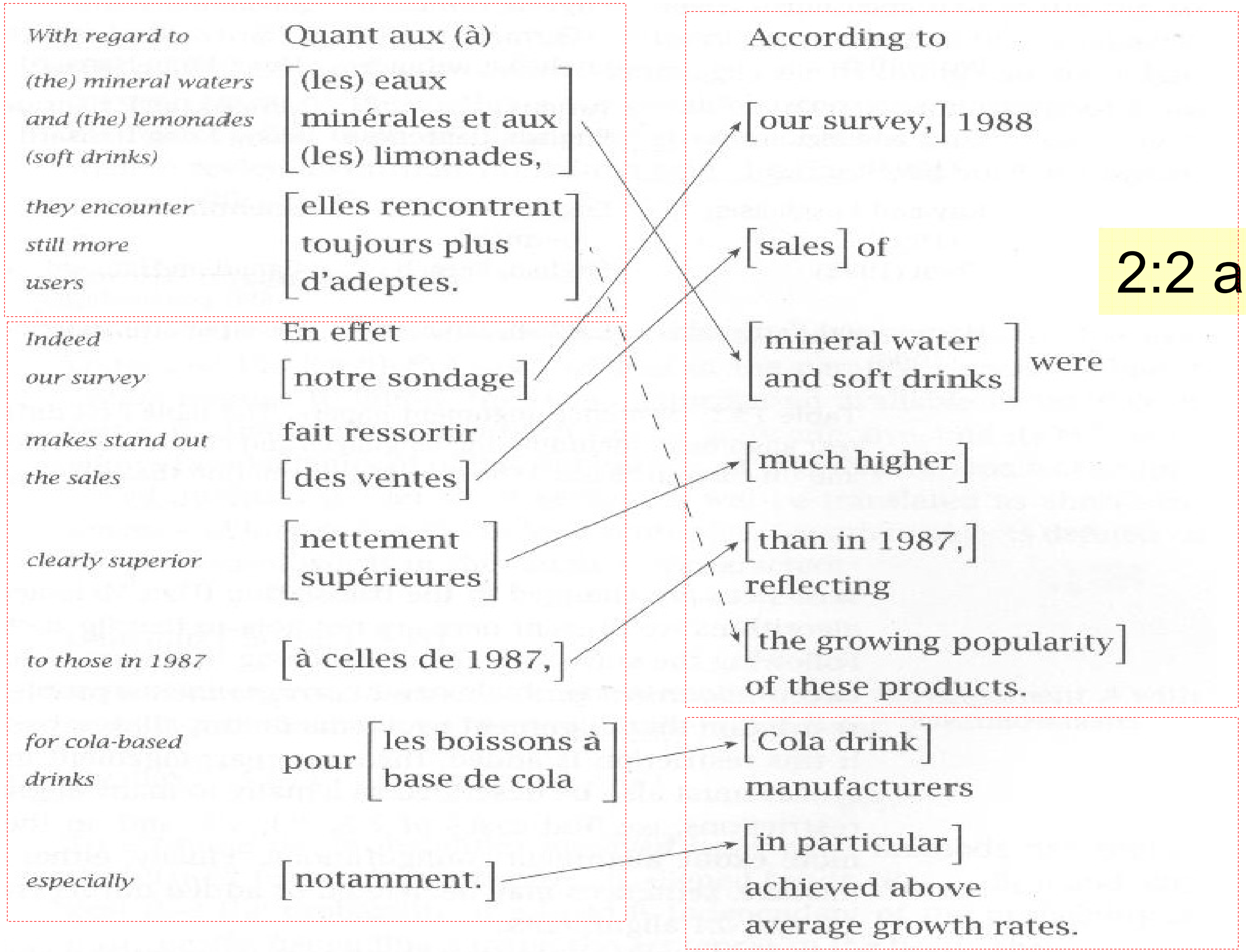
- Sources of Parallel texts or bitexts
  - Parliamentary proceedings (Hansards)
  - Newspapers and magazines
  - Religious and literary works

with less literal translation

- Two levels of alignment
  - **Gross large scale alignment**
    - Learn which paragraphs or sentences correspond to which paragraphs or sentences in another language
  - **Word alignment**
    - Learn which words tend to be translated by which words in another language
    - The necessary step for acquiring a bilingual dictionary

Orders of word or sentence might not be preserved.

Text Alignment: Example 1



With regard to	Quant aux (à)
(the) mineral waters	[(les) eaux
and (the) lemonades	minérales et aux
(soft drinks)	(les) limonades,]
they encounter	[elles rencontrent
still more	toujours plus
users	d'adeptes.]
Indeed	En effet
our survey	[notre sondage]
makes stand out	fait ressortir
the sales	[des ventes]
clearly superior	[nettement supérieures]
to those in 1987	[à celles de 1987,]
for cola-based drinks	pour [les boissons à base de cola]
especially	[notamment.]

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]

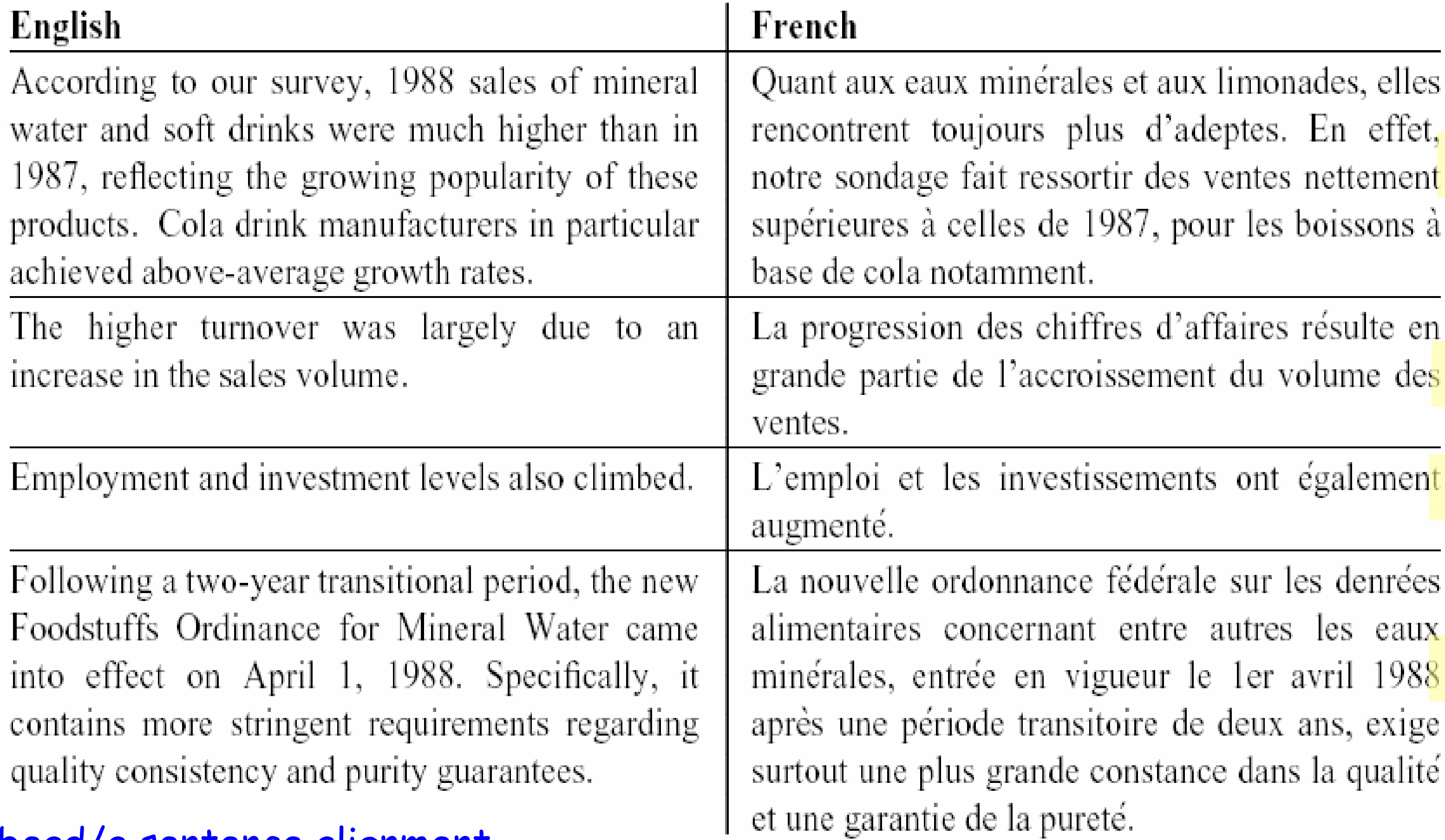
achieved above

average growth rates.

2:2 alignment

Figure 13.2 Alignment and correspondence. The middle and right columns show the French and English versions with arrows connecting parts that can be viewed as translations of each other. The italicized text in the left column is a fairly literal translation of the French text.

Text Alignment: Example 2



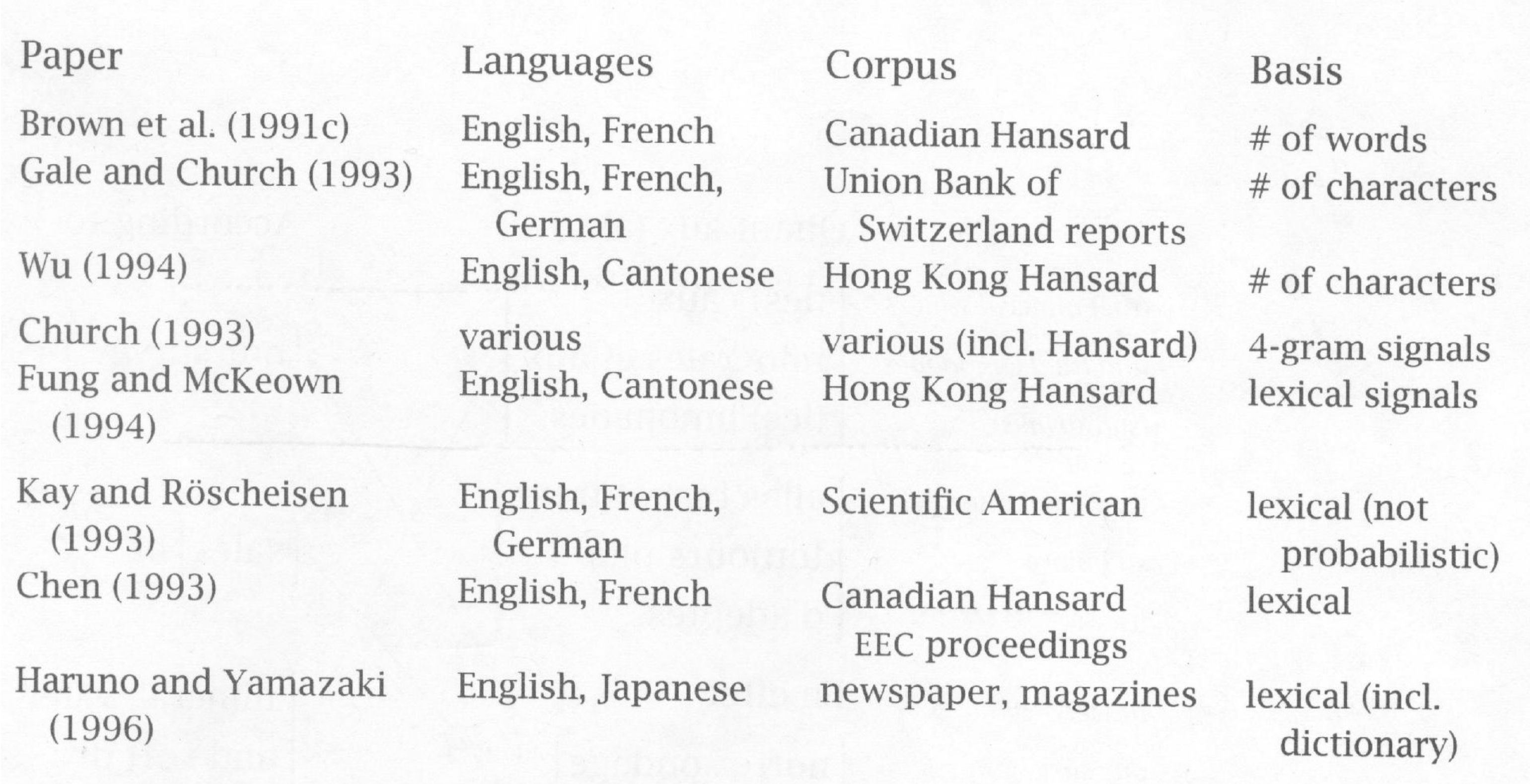
English	French	
According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates.	Quant aux eaux minérales et aux limonades, elles rencontrent 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.	2:2 alignment
The higher turnover was largely due to an increase in the sales volume.	La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes.	1:1 alignment
Employment and investment levels also climbed.	L'emploi et les investissements ont également augmenté.	1:1 alignment
Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.	La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.	2:1 alignment

a bead/a sentence alignment

Studies show that around 90% of alignments are 1:1 sentence alignment.

Sentence Alignment

- Crossing dependencies are not allowed here
  - Word ordering is preserved !
- Related work



Paper	Languages	Corpus	Basis
Brown et al. (1991c)	English, French	Canadian Hansard	# of words
Gale and Church (1993)	English, French, German	Union Bank of Switzerland reports	# of characters
Wu (1994)	English, Cantonese	Hong Kong Hansard	# of characters
Church (1993)	various	various (incl. Hansard)	4-gram signals
Fung and McKeown (1994)	English, Cantonese	Hong Kong Hansard	lexical signals
Kay and Röscheisen (1993)	English, French, German	Scientific American	lexical (not probabilistic)
Chen (1993)	English, French	Canadian Hansard EEC proceedings	lexical
Haruno and Yamazaki (1996)	English, Japanese	newspaper, magazines	lexical (incl. dictionary)

Sentence Alignment

- Length-based
- Lexical-guided
- Offset-based

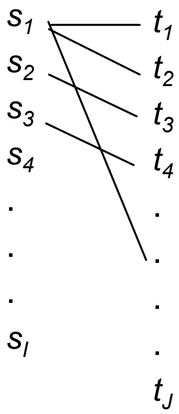
Sentence Alignment
Length-based method

- **Rationale**: the short sentences will be translated as short sentences and long sentences as long sentences
  - Length is defined as the number of words or the number of characters

- **Approach 1** (Gale & Church 1993)

Union Bank of Switzerland (UBS) corpus
: English, French, and German

  - **Assumptions**
    - The paragraph structure was clearly marked in the corpus, confusions are checked by hand
    - Lengths of sentences measured in characters
    - **Crossing dependences** are not handled here
      - The order of sentences are not changed in the translation



Ignore the rich information available in the text.

Sentence Alignment
Length-based method

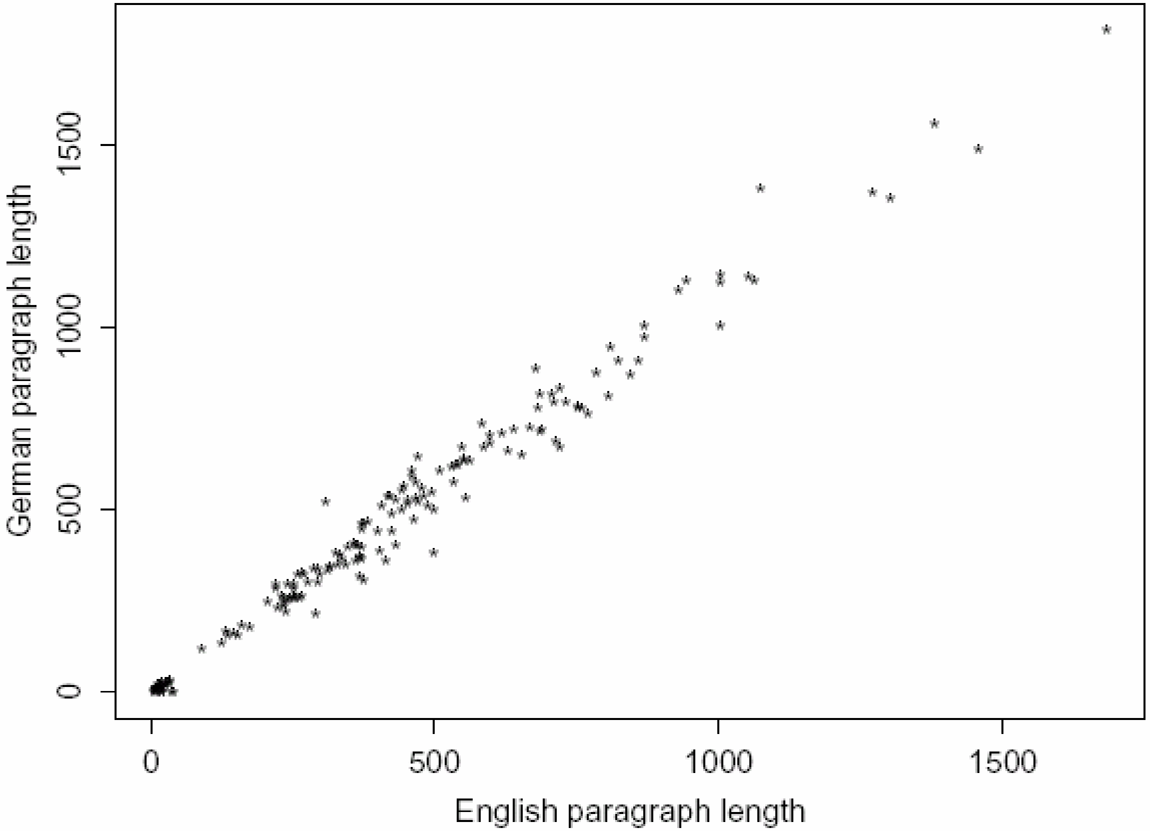


Figure 1. The horizontal axis shows the length of English paragraphs, while the vertical scale shows the lengths of the corresponding German paragraphs. Note that the correlation is quite large (.991).

Most cases are 1:1 alignments.

Sentence Alignment

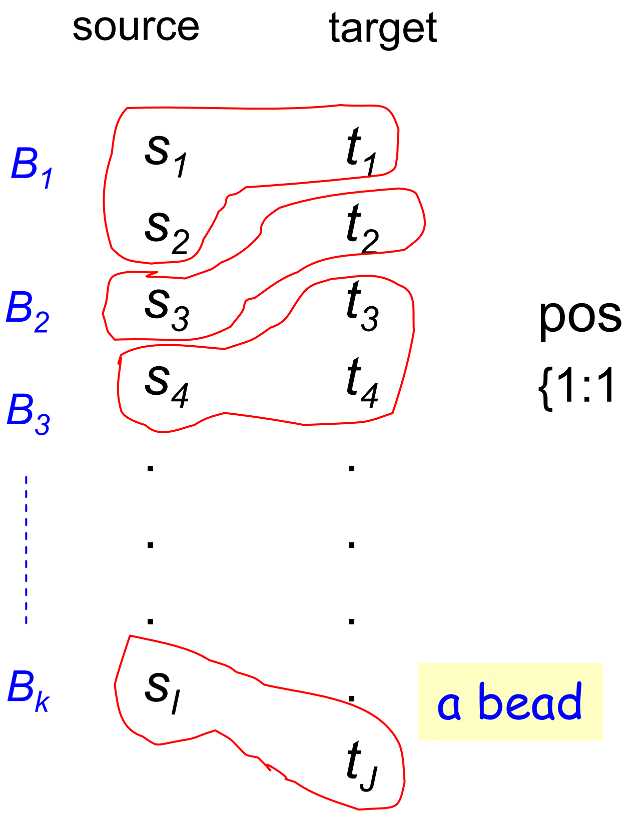
Length-based method

Source

$$S = s_1 s_2 \cdots s_I$$

$$T = t_1 t_2 \cdots t_J$$

Target



possible alignments:

{1:1, 1:0, 0:1, 2:1,1:2, 2:2,…}

probability independence between beads

$$\operatorname*{argmax}_{A} P(A|S,T) = \operatorname*{argmax}_{A} \boxed{P(A,S,T) \quad (\approx \prod_{k=1}^{K} P(B_k))}$$

$$\text{where} \quad A = (B_1, B_2, \ldots, B_k)$$

Sentence Alignment
Length-based method

- Dynamic Programming
  - The cost function (Distance Measure)

Bayes' Law

$$\operatorname{cost}(\alpha \text{ align } l_1, l_2) = -\log P\left(\alpha \text{ align} \,\middle|\, \delta\left(l_1, l_2, \mu, s^2\right)\right)$$

$$\approx -\log\left[P(\alpha \text{ align})P\left(\delta\left(l_1, l_2, \mu, s^2\right) \,\middle|\, \alpha \text{ align}\right)\right]$$

$-\log P(B_k)$

$$\delta\left(l_1, l_2, \mu, s^2\right) = (l_2 - l_1\mu) \Big/ \sqrt{l_1 s^2}$$

$\delta(\cdot)$ is a distance measure which forms a normal distribution

Ratio of texts in two languages $\frac{L_2}{L_1} = \mu$

square difference of two paragraphs

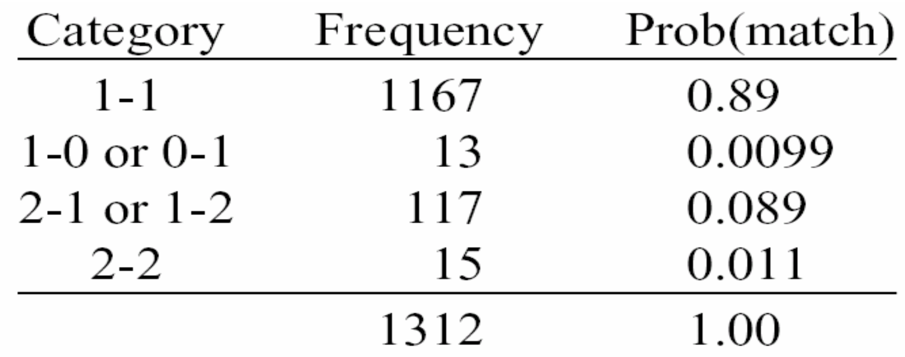
- Sentence is the unit of alignment
- Statistically modeling of character lengths

$$P\left(\delta\left(l_1, l_2, \mu, s^2\right) \,\middle|\, \alpha \text{ align}\right) = 2\left(1 - \boxed{prob\left(\left|\delta\right|\right)}\right)$$

The prob. distribution of standard normal distribution

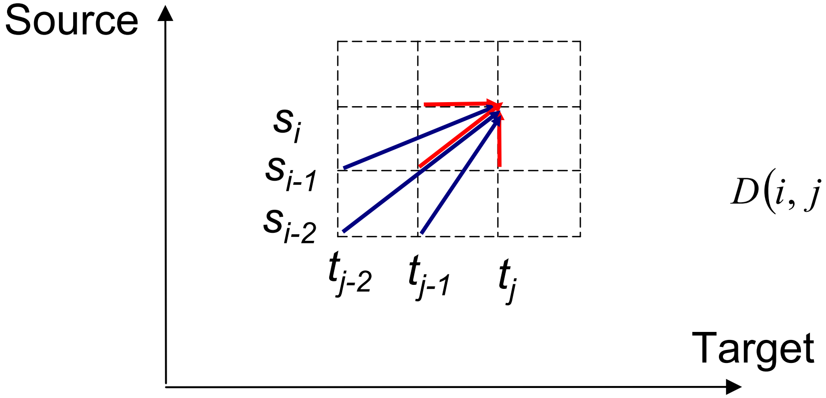
Sentence Alignment
Length-based method

- The priori probability



Category	Frequency	Prob(match)
1-1	1167	0.89
1-0 or 0-1	13	0.0099
2-1 or 1-2	117	0.089
2-2	15	0.011
	1312	1.00

Or $P(\alpha \text{ align})$



$$D(i,j)=\begin{cases} D(i,j-1)+\text{cost}\left(0:1\text{ align }\phi,t_j\right)\\ D(i-1,j)+\text{cost}\left(1:0\text{ align }s_i,\phi\right)\\ D(i-1,j-1)+\text{cost}\left(1:1\text{ align }s_i,t_j\right)\\ D(i-1,j-2)+\text{cost}\left(1:2\text{ align }s_i,t_{j-1},t_j\right)\\ D(i-2,j-1)+\text{cost}\left(2:1\text{ align }s_{i-1},s_i,t_j\right)\\ D(i-2,j-2)+\text{cost}\left(2:2\text{ align }s_{i-1},s_i,t_{j-1},t_j\right)\end{cases}$$

Sentence Alignment
Length-based method

– A simple example

L_1 alignment 1

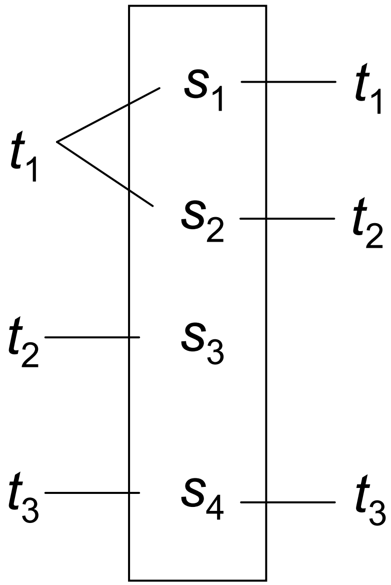
cost(align(s_1, s_2, t_1))

+

cost(align(s_3, t_2))

+

cost(align(s_4, t_3))



L_1 alignment 2

cost(align(s_1, t_1))

+

cost(align(s_2, t_2))

+

cost(align(s_3,Ø))

+

cost(align(s_4, t_3))

Sentence Alignment
Length-based method

– The experimental results

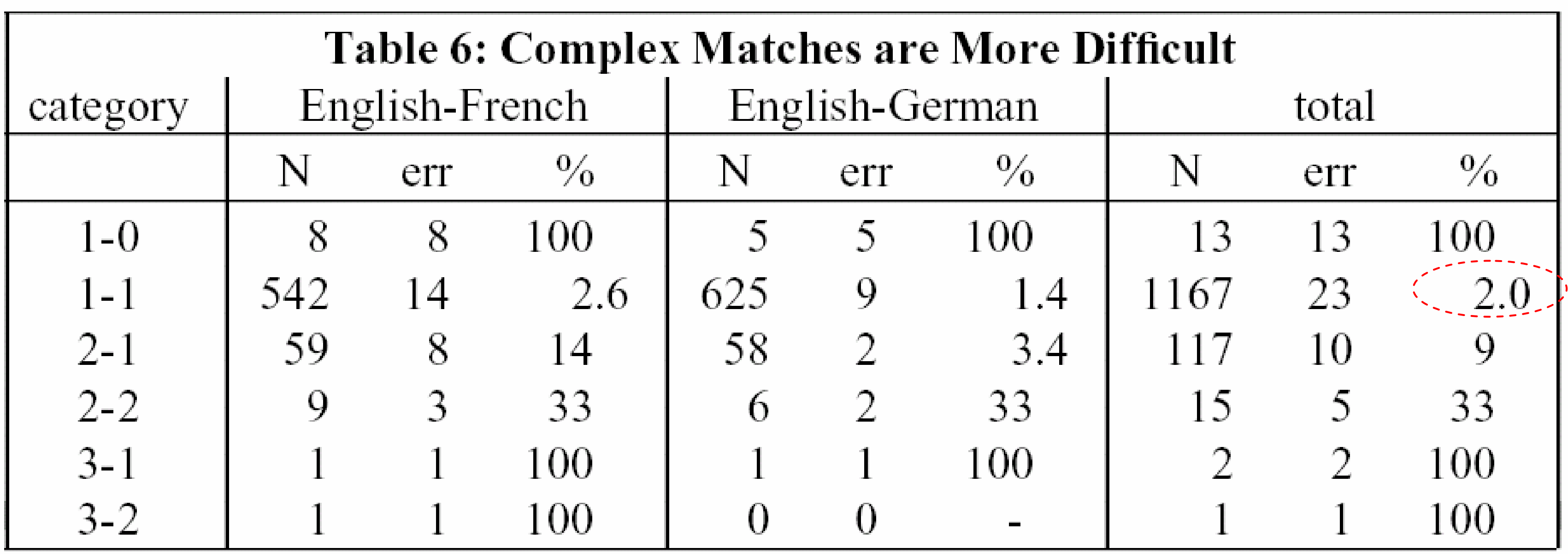


Table 6: Complex Matches are More Difficult

category	English-French			English-German			total		
	N	err	%	N	err	%	N	err	%
1-0	8	8	100	5	5	100	13	13	100
1-1	542	14	2.6	625	9	1.4	1167	23	2.0
2-1	59	8	14	58	2	3.4	117	10	9
2-2	9	3	33	6	2	33	15	5	33
3-1	1	1	100	1	1	100	2	2	100
3-2	1	1	100	0	0	-	1	1	100

Sentence Alignment

Length-based method

- 4% error rate was achieved
- **Problems**:
  - Can not handle noisy and imperfect input
    - E.g., OCR output or file containing unknown markup conventions
    - Finding paragraph or sentence boundaries is difficult
    - **Solution**: just align text (position) offsets in two parallel texts (Church 1993)
  - Questionable for languages with few cognates or different writing systems
    - E.g., English ⟷ Chinese

eastern European languages ⟷ Asian languages

Sentence Alignment
Length-based method

- **Approach 2** (Brown 1991)
  - Compare sentence length in words rather than characters
    - However, variance in number of words us greater than that of characters
  - EM training for the model parameters

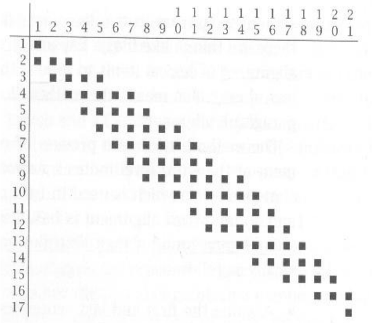
- **Approach 3** (Wu 1994)
  - Apply the method of Gale and Church(1993) to a corpus of parallel English and Cantonese text
  - Also explore the use of lexical cues

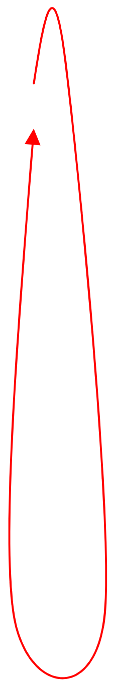
Sentence Alignment
Lexical method

- **Rationale**: the lexical information gives a lot of confirmation of alignments
  - Use a partial alignment of lexical items to induce the sentence alignment
  - That is, a partial alignment at the word level induces a maximum likelihood at the sentence level
  - The result of the sentence alignment can be in turn to refine the word level alignment

Sentence Alignment
Lexical method

- Approach 1 (Kay and Röscheisen 1993)
  - First assume the first and last sentences of the text were align as the initial anchors
  - Form an envelope of possible alignments
    - Alignments excluded when sentences across anchors or their respective distance from an anchor differ greatly
  - Choose word pairs their distributions are similar in most of the sentences
  - Find pairs of source and target sentences which contain many possible lexical correspondences
    - The most reliable of pairs are used to induce a set of partial alignment (add to the list of anchors)





Iterations

Sentence Alignment

Lexical method

- Approach 1
  - Experiments
    - On Scientific American articles
      - 96% coverage achieved after 4 iterations, the reminders is 1:0 and 0:1 matches
    - On 1000 Hansard sentences
      - Only 7 errors (5 of them are due to the error of sentence boundary detection) were found after 5 iterations
  - Problem
    - If a large text is accompanied with only endpoints for anchors, the pillow must be set to large enough, or the correct alignments will be lost
      - Pillow is treated as a constraint

Sentence Alignment
Lexical method

- Approach 2 (Chen 1993)
  - Sentence alignment is done by constructing a simple word-to-word alignment
  - Best alignment is achieved by maximizing the likelihood of the corpus given the translation model
  - Like the method proposed by Gale and Church(1993), except that a translation model is used to estimate the cost of a certain alignment

$$\arg\max_{A} P(A, S, T) \approx \prod_{k=1}^{K} P(B_k)$$

The translation model

$$\begin{aligned} -\log P(B_k) &= \text{cost}(\alpha \text{ align } l_1, l_2) \\ &\approx -\log\left[P(\alpha \text{ align}) P(T(l_1, l_2) \mid \alpha \text{ align})\right] \end{aligned}$$

Sentence Alignment
Lexical method

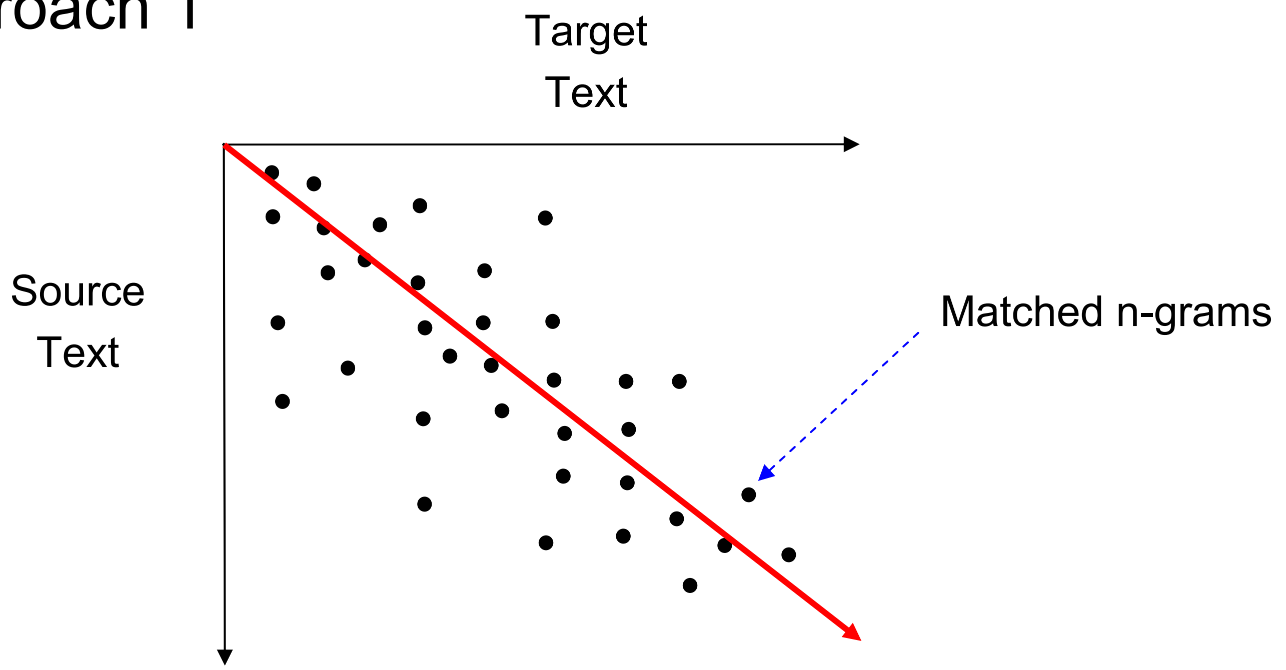
- Approach 3 (Haruno and Yamazaki, 1996)
  - Function words are left out and only content words are used for lexical matching
  - Part-of-speech taggers are needed
  - For short texts, an on-line dictionary is used instead of the finding of word correspondences adopted by Kay and Röscheisen (1993)

Offset Alignment

- Perspective
  - Do not attempt to align beads of sentences but just align position offsets in two parallel texts
  - Avoid the influence of noises or confusions in texts
    - Can alleviate the problems caused by the absence of sentence markups
- Approach 1: (Church 1993)
  - Induce an alignment by cognates, proper nouns, numbers, etc.
    - **Cognate words**: words similar across languages
    - **Cognate words** share ample supply of identical character sequences between source and target languages
  - Use DP to find a alignment for the occurrence of matched character 4-grams along the diagonal line

Offset Alignment

- Approach 1



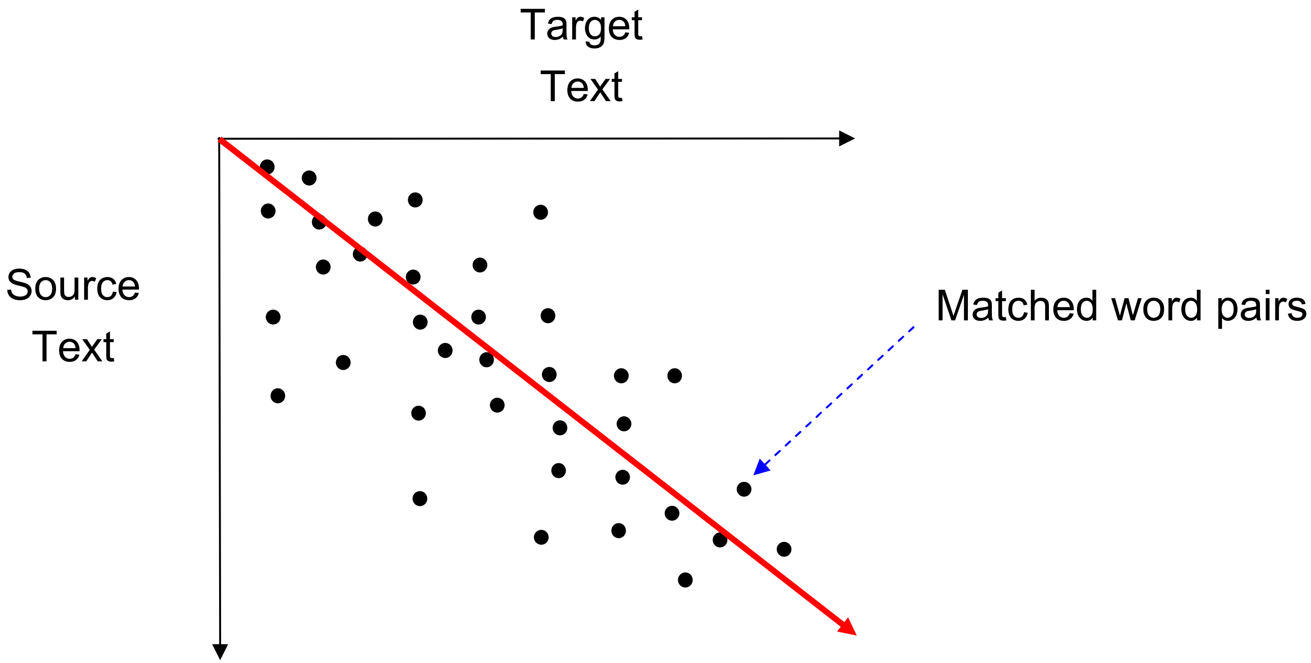
  - Problem
    - Fail completely when language with different character sets (English ←→Chinese)

Offset Alignment

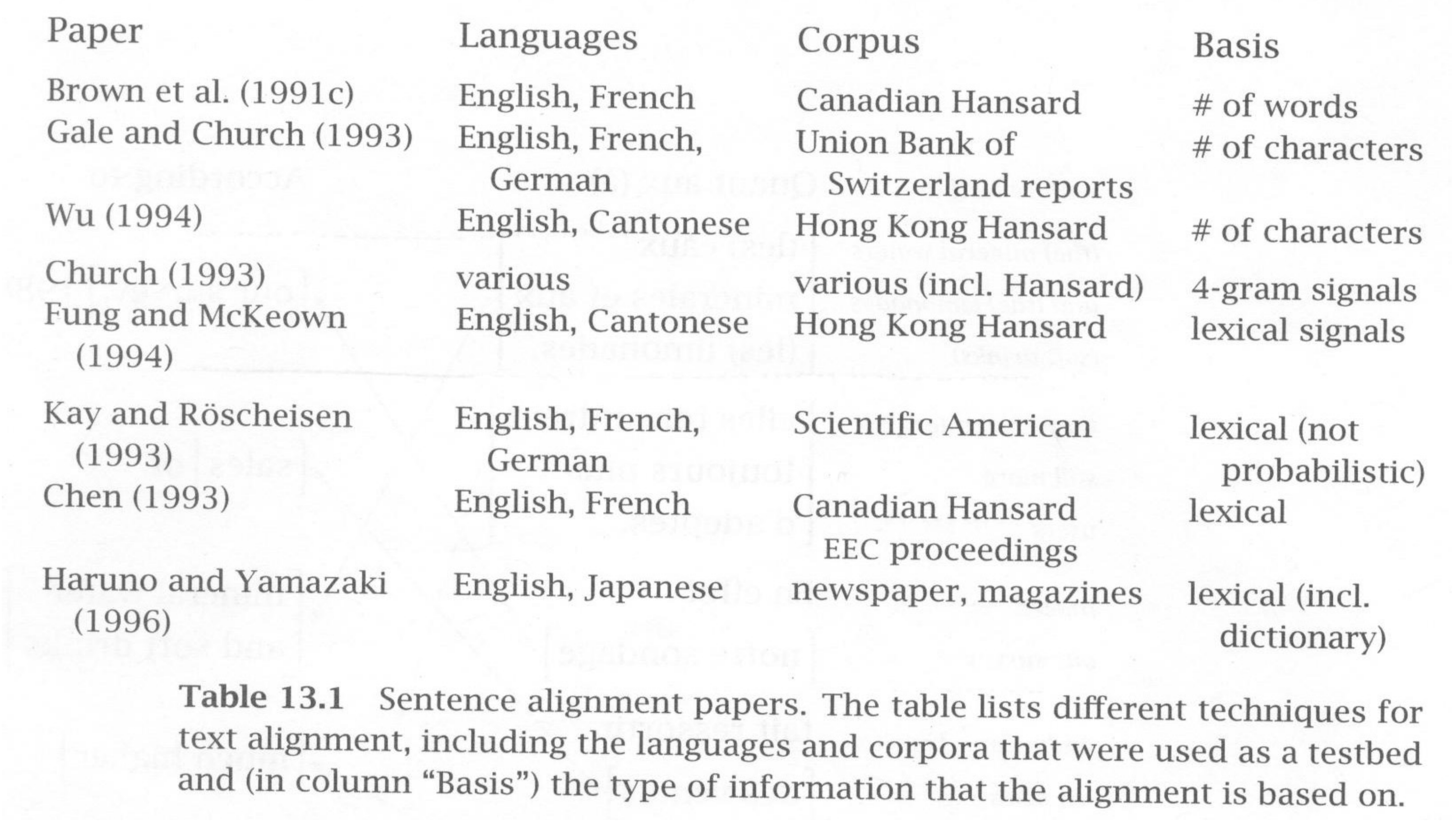
- Approach 2: (Fung and McKeown 1993)
  - Two-sage processing
  - First stage (to infer a small bilingual dictionary)
    - For each word a signal is produced, as an arrive vector of integer number of words between each occurrence
      - E.g., word appears in offsets (1, 263, 267, 519) has an arrival vector (262,4,252)
    - Perform Dynamic Time Warping to match the arrival vectors of two English and Cantonese words to determine the similarity relations
    - Pairs of an English word and Cantonese word with very similar signals are retained in the dictionary
  - Properties
    - Genuinely language independent
    - Sensitive to lexical content

Offset Alignment

- Approach 2: (Fung and McKeown 1993)
  - Second stage
    - Use DP to find a alignment for the occurrence of strongly-related word pairs along the diagonal line



Sentence/Offset Alignment: Summary



Paper	Languages	Corpus	Basis
Brown et al. (1991c)	English, French	Canadian Hansard	# of words
Gale and Church (1993)	English, French, German	Union Bank of Switzerland reports	# of characters
Wu (1994)	English, Cantonese	Hong Kong Hansard	# of characters
Church (1993)	various	various (incl. Hansard)	4-gram signals
Fung and McKeown (1994)	English, Cantonese	Hong Kong Hansard	lexical signals
Kay and Röscheisen (1993)	English, French, German	Scientific American	lexical (not probabilistic)
Chen (1993)	English, French	Canadian Hansard EEC proceedings	lexical
Haruno and Yamazaki (1996)	English, Japanese	newspaper, magazines	lexical (incl. dictionary)

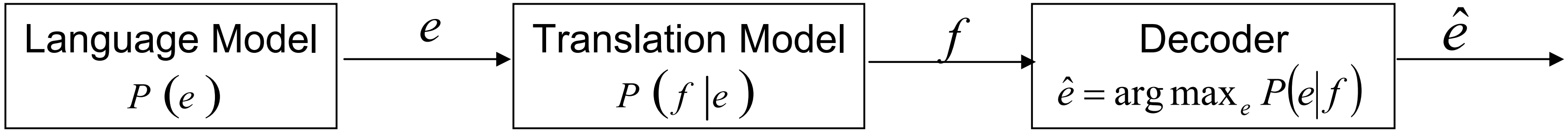
Table 13.1 Sentence alignment papers. The table lists different techniques for text alignment, including the languages and corpora that were used as a testbed and (in column "Basis") the type of information that the alignment is based on.

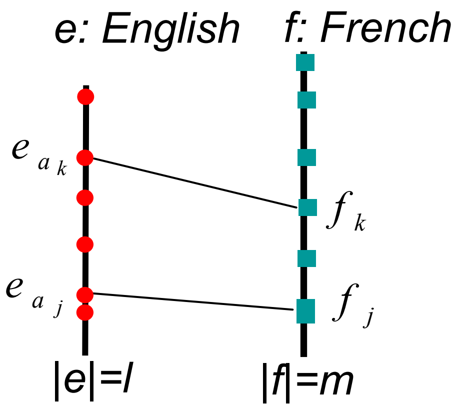
Word Alignment

- The sentence/offset alignment can be extended to a word alignment
- Some criteria are then used to select aligned word pairs to include them into the bilingual dictionary
  - Frequency of word correspondences
  - Association measures
  - ….

Statistical Machine Translation

- The noisy channel model





  - Translation in sentence level
  - **Assumptions**:
    - An English word can be aligned with multiple French words while each French word is aligned with at most English word
    - Independence of the individual word-to-word translations

Statistical Machine Translation

- Three important components involved
  - Language model
    - Give the probability *p*(*e*)
  - Translation model

$$P(f|e)=\frac{1}{Z}\underbrace{\sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}}_{\substack{\text{all possible alignments}\\ \text{(the English word that a French word } f_j \text{ is aligned with)}}}\prod_{j=0}^{m}P\left(f_j|e_{a_j}\right)$$

normalization constant

translation probability

  - Decoder

$$\hat{e}=\arg\max_{e} P(e|f)=\arg\max_{e}\frac{p(e)p(f|e)}{p(f)}=p(e)p(f|e)$$

Statistical Machine Translation

- EM Training
  - E-step (Expectation)

$$Z_{w_f, w_e} = \sum_{(e,f) \text{s.t. } w_e \in e, w_f \in f} P(w_f | w_e)$$

$$\uparrow$$

Number of times that w_e occurred in the English sentences while w_f in the corresponding French sentences

  - M-step (Maximization)

$$P(w_f | w_e) = \frac{Z_{w_f, w_e}}{\sum_v Z_{v, w_e}}$$