Sentence Alignment

Multilingual NLP

Guillaume Wisniewski
guillaume.wisniewski@linguist.univ-paris-diderot.fr
October 2019

Université de Paris & LLF

Sentence

Definition

- meaningful grammatical structure ⊕ self-contained
- express some kind of statement / request / command / ...

Why considering sentences ?

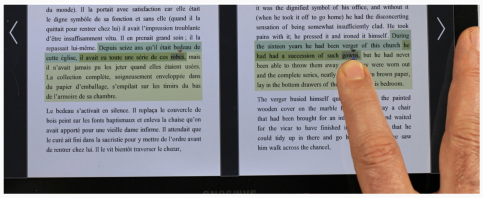
- meaningful unit ⊕ short enough to be processed efficiently
- many applications
- sentence-aligned bitexts = most important resource for MT ⊕ computer-aided translation

Disclaimer

- assumption : documents have been segmented into sentences (and tokenized ?)
- not that easy ! especially in a multilingual context
- has a large impact on the quality of sentence alignment even manual alignment
  - ↪ (Simard 1998) most annotator disagreements are caused by questions of sentence segmentation rather than questions of translational equivalence

We will not talk about this problem !

Sentence alignment



Transread bilingual reader
https://transread.limsi.fr

Mapping of adjacent *source* sentences to their corresponding *target* sentences in a bitext

Formally

Definition

- given a bitext with a set of M source sentences $\mathbf{E}_1^M = (E_1, ..., E_M)$ and N target sentences $\mathbf{F}_1^N = (F_1, ..., F_N)$
- find the corresponding (i.e. same meaning) sentence groups
- sentence groups = consecutive sentences (possibly none)

Example

E_1	I am giving a talk.	link 1	Je fais une présentation.	F_1
E_2	I am really hungry.	link 2	J'ai très faim.	F_2
E_3	I want to eat something.	link 3	Je veux manger.	F_3

Formally

Definition

- given a bitext with a set of M source sentences $\mathbf{E}_1^M = (E_1, ..., E_M)$ and N target sentences $\mathbf{F}_1^N = (F_1, ..., F_N)$
- find the corresponding (i.e. same meaning) sentence groups
- sentence groups = consecutive sentences (possibly none)

Example

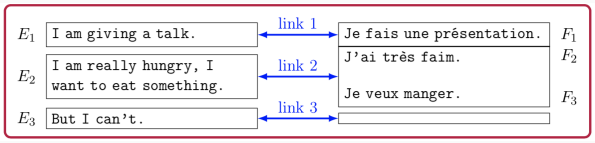
E_1	I am giving a talk.	link 1	Je fais une présentation.	F_1
E_2	I am really hungry, I want to eat something.	link 2	J'ai très faim. Je veux manger.	F_2 F_3

Formally

Definition

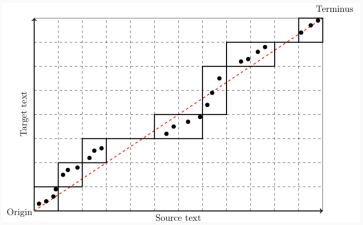
- given a bitext with a set of M source sentences $\mathbf{E}_1^M = (E_1, ..., E_M)$ and N target sentences $\mathbf{F}_1^N = (F_1, ..., F_N)$
- find the corresponding (i.e. same meaning) sentence groups
- sentence groups = consecutive sentences (possibly none)

Example



Bitext Space

- visualization of bitext alignment (Melamed 1999)



↪ source words on the x-axis
↪ dashed lines identify sentence boundaries
↪ sub-rectangle = alignment links
↪ horizontal line (resp. vertical) = null link with target (resp. source)

Number of possible alignments ?

- given a source text of N sentences...
- ... and its translation in M sentences...
- how many possible alignments are there ?

And the answer is...

- too many !

And the answer is...

- too many !
- number of possible segmentation of the source ? 2^N

And the answer is...

- too many !
- number of possible segmentation of the source ? 2^N
- number of possible 1 :1 alignments $\min(M, N)!$

And the answer is...

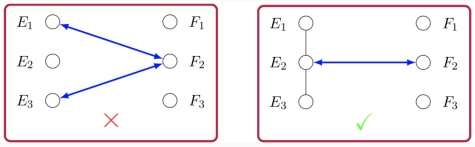
- too many !
- number of possible segmentation of the source ? 2^N
- number of possible 1 :1 alignments $\min(M, N)!$
- no need to go further !

Simplifying hypothesis

- motivated by the task...
- ...and our need to reduce the search space (computational simplification)

1st Assumption

- if E_i and E_{i+2} are inside a link, so is E_{i+1}

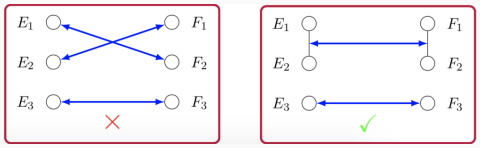


Simplifying hypothesis

- motivated by the task...
- ...and our need to reduce the search space (computational simplification)

2nd Assumptions

- alignment links are monotonic



Simplifying hypothesis

- motivated by the task...
- ...and our need to reduce the search space (computational simplification)

3rd Assumptions

- restriction on the number of sentences in a group
- e.g. : considering only `1:1`, `0:1` or `1:0` alignments.
- in general : `n:m` with $n, m \in \{0, 1, 2\}^2$

Typology of approaches

1. length-based
2. lexical-based
3. combined

Length-Based Methods

A brilliant idea...

Short sentences are translated by short sentences.
Long sentences are translated by long sentences.

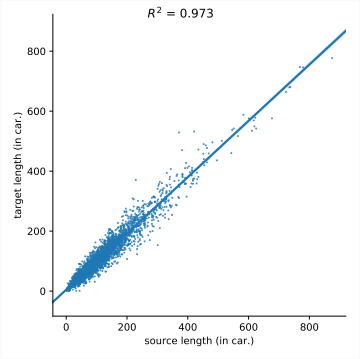
(Gale et Kenneth W. Church 1993)

↪ really true ?

↪ in all contexts (language pairs ? genre of documents ?)

Correlation between sentence lengths (1)

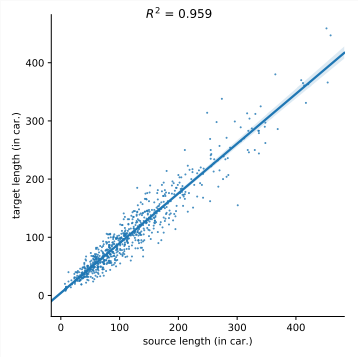
French-English (Novel)



estimated on 11 French Novels hand-aligned
https://transread.limsi.fr/Resources/ReferenceSentenceAlignment.tgz

Correlation between sentence lengths (2)

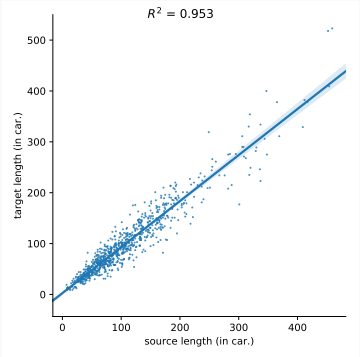
Spanish-English (Novel)



estimated on 1 novel hand-aligned
https://transread.limsi.fr/Resources/ReferenceSentenceAlignment.tgz

Correlation between sentence lengths (2)

Greek-English (Novel)



estimated on 1 novel hand-aligned
https://transread.limsi.fr/Resources/ReferenceSentenceAlignment.tgz

Alignment Model (1) : search space

Definition of the search space / alignments
4 kinds of alignments :

↪ 1:1 : direct translation

↪ 1:0 or 0:1 : insertion or deletion

↪ 2:1 or 1:2 : expansion or contraction

↪ 2:2 : swap or merge

Alignment Model (1) : search space

Definition of the search space / alignments
4 kinds of alignments :

↪ 1:1 : direct translation 89%

↪ 1:0 or 0:1 : insertion or deletion 0.99%

↪ 2:1 or 1:2 : expansion or contraction 8.9%

↪ 2:2 : swap or merge 1.1%

⇒ Most frequent kind of alignments

Alignment Model (2) : search space

- aligning two documents ⇒ find the 'best' sequence of operations
- in a combinatorial space

Alignment Model (2) : search space

- aligning two documents ⇒ find the 'best' sequence of operations
- in a combinatorial space

⇒ dynamic programming

↪ closely related to compute the edit distance

Alignment Model (3)

Cost function

$$p(\text{match}|l_1, l_2) \propto p(\text{match}) \times p(l_1, l_2|\text{match})$$

↪ probability that a sentence of length l_1 is aligned to a sentence of length l_2

↪ decision based solely on the length of the sentence

And now ?

- How can we model $p(l_1, l_2|\text{match})$
- (Gale et Kenneth W. Church 1993) 'main' idea :
  - ↪ $l_1 - l_2$ follows a normal distribution $\mathcal{N}(\mu, \sigma)$ when the sentences are aligned
  - ↪ how likely is it to observe a length difference as large as $l_1 - l_2$?

Technical details

Let us define :

$$\delta(l_i, l_j) = \frac{l_j - c \times l_i}{\sqrt{\frac{1}{2} \times (l_j + c \times l_i) \times s^2}}$$

where c and s^2 are chosen so that $\delta(l_i, l_j) \sim \mathcal{N}(0, 1)$ (normalization)

Technical details

The probability to observe a value larger than δ assuming that $\delta \sim \mathcal{N}(0, 1)$ is

$$p\left(|X| \geq |\delta\left(l_i, l_j\right)|\right) = 2 \cdot \left(1 - p(X < |\delta(l_i, l_j)|)\right)$$

with :

$$p(X < |\delta(l_i, l_j)|) = \frac{1}{\sqrt{2 \cdot \pi}} \times \int_{-\infty}^{\delta} e^{\frac{-z}{2}} \cdot dz$$

↪ value can be found in standard tables / directly computed by many libraries

Technical details

How to choose c and s ?

↪ estimated on a dataset (only need a parallel corpus no alignment)

↪ constant across languages and corpora : $c = 1$ and $s^2 = 6.8$

At the end

Given four sequences (x_1, x_2, y_1, y_2), we can define the following costs :

- $d(x_1, y_1, 0, 0)$: the cost of aligning x_1 to y_i
- $d(x_1, 0, 0, 0)$: the cost of deleting x_1
- $d(0, y_1, 0, 0)$: the cost of deleting y_1
- $d(x_1, y_1, x_2, 0)$: the cost of contracting x_1, x_2 into y_2
- $d(x_1, y_1, 0, y_2)$: the cost of expanding x_1 to y_1, y_2
- $d(x_1, y_1, x_2, y_2)$: the cost of aligning the 4 sentences together

Dynamic Program

Initialization :

$$D(0,0) = 0$$

Recursion :

$$D(i,j) = \min \begin{cases} D(i,j-1) & +d(0,t_j,0,0) \quad \text{(insertion)} \\ D(i-1,j) & +d(s_i,0,0,0,) \quad \text{(suppression)} \\ D(i-1,j-1) & +d(s_i,t_j,0,0) \quad \text{(direct translation)} \\ D(i-1,j-2) & +d(s_i,t_j,0,t_{j-1}) \quad \text{(expansion)} \\ D(i-2,j-1) & +d(s_i,t_j,s_{i-1},0) \quad \text{(contraction)} \\ D(i-2,j-2) & +d(s_i,t_j,s_{i-1},t_{j-1}) \quad \text{(swap/marge)} \end{cases}$$

$\Rightarrow$ exactly as the DP for computing an edit distance

Performance

Pros

- fast
- good performance
- has been applied to a wide array of languages (even if the assumption on δ are no longer true)

Cons

- can not be applied on noisy bitexts [Le *et al.*, 2010]
  - different information on both side of the bitext
  - missing/added sentence or paragraph
- rely on a single information $\Rightarrow$ not robust $\oplus$ risk of error propagation

Lexical Matching Approaches

Intuition

- corresponding sentences contain many lexical translational equivalent
- monotonicity assumption $\Rightarrow$ alignment in the neighborhood of the bitext diagonal

$\Downarrow$

Two steps approach

1. identify anchor point candidates
2. align sentences according to these anchors

Identify anchors

Surface comparison

- identify common words (number, punctuations, ...)
- cognates (highly similar words that can be identified by string matching techniques)

Dictionaries

- identify common translations using a lexicon

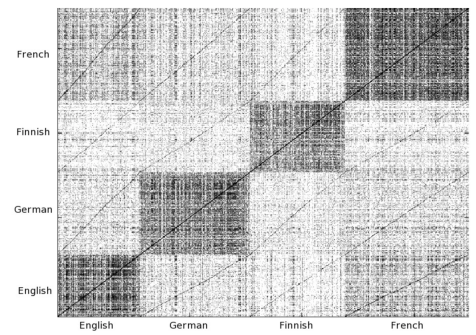
$\Rightarrow$ chicken-egg problem

Common 4-grams across languages

Dot-plot method (Kenneth Ward Church et Helfman 1993)

↪ KDE documentation in 4 languages (400 sentences)

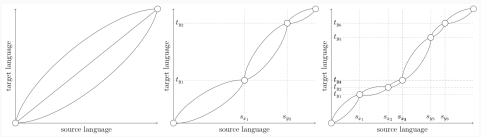
↪ a dot for every common 4-gram



Matching sentences (1)

Geometric Interpretation (Melamed 1999)

- find a 'smooth' path in the bitext space
- iterative refinement



Matching sentences (2)

Dynamic Programming

Lexical Approaches : Pros & Cons

- fragile :
  - ↪ decisions are never questioned
  - ↪ anchor should be used a soft constraints
- rely on the existence of a lexicon
  - ↪ available for all language pairs ?
  - ↪ not as general as length-based approaches
- able to detect missing part in translations

Combining lexical & length information

The Champollion aligner (Ma 2006)

Three main ideas :

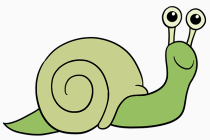
1. use lexical match
2. weight matches using tf-idf
   - less frequent translation term pairs have much stronger evidence for two segments to be aligned
3. penalize matches according to difference in lengths

Performance

Alignment quality

- Chinese-English ; $\simeq 5,000$ sentences
- Recall $= 96.9\%$ Precision $= 97.0\%$

But...



- `Champollion` is slow
- many look-ups in dictionaries
- match in $\mathcal{O}\left(n^2\right)$

The sentence aligner of (Moore 2002)

The challenges

- use both lexical and length-based information
- no need for a lexicon

How ?

- A two-pass approach :
  1. align sentences using a length based model
  2. automatically extract a lexicon & re-align the corpus with this additional information
     - ↪ use a word-alignment model (adaptation of IBM 1 model)

Yasa (Lamraoui et Langlais 2013)

Score of aligning a bisegment :

$$S_{\text{cognate}} = -\lambda_1 \cdot \left(\left[c \log \frac{p_T}{p_R} \right] - \left[(n-c) c \log \frac{1-p_T}{1-p_R} \right] \right)$$

$$S_{\text{length}} = -\lambda_2 \cdot P(\delta|\text{match})$$

$$S_{\text{prior}} = -\lambda_3 \cdot P(\text{match})$$

first term : likelihood ratio → how likely a pairing involving n words on average share c cognates under the assumption that the sentences are in translation or not.

- ↪ final criterion : linear combination of 3 criteria
- ↪ weights can be 'learned' on a development corpus (using a derivative free method such as minimization simplex technique of (Nelder et Mead 1965))

Conclusions

Performance Comparison

- bilingual parliament proceedings, manual, ...
- $F_1 \simeq 95\%$

Literary works

	(Melamed 1999)	(Moore 2002)	(Lamraoui et Langlais 2013)
min	53.5	57.4	59.9
max	92.8	91.5	95.6
mean	79.6	74.9	79.1

Is the task solved ?

- As usual :
- why are we aligning sentences ?
  1. to train MT system
     - ↪ usually only considers high quality 1 :1 alignment
     - ↪ discard a lot of data
  2. to analyze the translation process, cross-lingual information retrieval, ...
     - ↪ need to align the complete text
     - ↪ not there yet

Références

- Church, Kenneth Ward et Jonathan Isaac Helfman (1993). "Dotplot : A Program for Exploring Self-Similarity in Millions of Lines of Text and Code". In : *Journal of Computational and Graphical Statistics* 2.2, p. 153–174. ISSN : 10618600. URL : http://www.jstor.org/stable/1390697.
- Gale, William A. et Kenneth W. Church (1993). "A Program for Aligning Sentences in Bilingual Corpora". In : *Comput. Linguist.* 19.1, p. 75–102. ISSN : 0891-2017. URL : http://dl.acm.org/citation.cfm?id=972450.972455.

- Lamraoui, Fethi et Philippe Langlais (2013). "Yet Another Fast, Robust and Open Source Sentence Aligner. Time to Reconsider Sentence Alignment ?". In : *XIV Machine Translation Summit*. Nice, France.
- Ma, Xiaoyi (2006). "Champollion : A Robust Parallel Text Sentence Aligner". In : *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy : European Language Resources Association (ELRA). URL : http://www.lrec-conf.org/proceedings/lrec2006/pdf/746_pdf.pdf.
- Melamed, I. Dan (1999). "Bitext Maps and Alignment via Pattern Recognition". In : *Comput. Linguist.* 25.1, p. 107–130. ISSN : 0891-2017. URL : http://dl.acm.org/citation.cfm?id=973215.973218.

- Moore, Robert C. (2002). "Fast and Accurate Sentence Alignment of Bilingual Corpora". In : *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation : From Research to Real Users*. AMTA '02. London, UK, UK : Springer-Verlag, p. 135–144. ISBN : 3-540-44282-0. URL : http://dl.acm.org/citation.cfm?id=648181.749407.
- Nelder, J. A. et R. Mead (1965). "A Simplex Method for Function Minimization". In : *Computer Journal* 7, p. 308–313.
- Simard, Michel (1998). "The BAF : A Corpus of English-French Bitext". In : *Proceedings of LREC 98*. Granada, Spain.