A Hybrid Word Alignment Approach to Improve Translation Lexicons with Compound Words and Idiomatic Expressions

(1) CEA–LIST/ Vision and Content Engineering Laboratory

and

(2) Softissimo

France

Nasredine Semmar (1), Christophe Servan (1), Gaël de Chalendar (1)

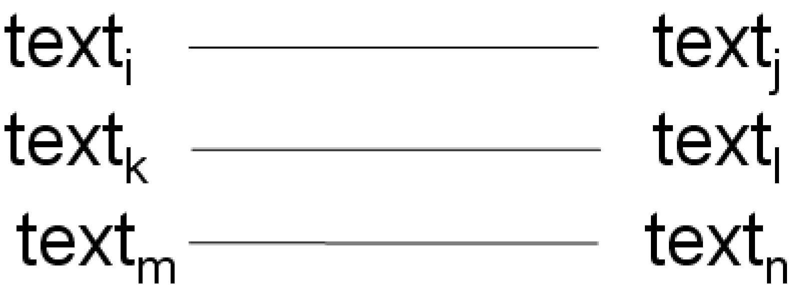
Benoît Le Ny (2), Jean-Jacques Bouzaglou (2)

Outline

- Parallel Texts
- Construction of Translation Lexicons
- Sentence Alignment
  - Sentence Alignement Process
  - Cross-language Information Retrieval
  - Multilingual Linguistic Analysis
- Word Alignment
  - Single-Word Alignment
  - Compound-Word Alignment
  - Collocation Alignment
- Experimental Results
- Conclusion and Future Work

Parallel Texts

- A parallel corpus is a collection of bi-texts



Source language **Target language**

text$_i$ — text$_j$

text$_k$ — text$_l$

text$_m$ — text$_n$

- ARCADE II corpus
  - **JOC (Official Journal of the European Community):** 1 million words in English, French, German, Italian and Spanish aligned at the sentence level
  - **MD (Le Monde Diplomatique):** 150 Arabic texts aligned to French at the sentence level, 50 aligned text pairs with French as pivot language for Russian, Chinese, Japanese, Greek and Persian

Outline

- Parallel Texts
- Construction of Translation Lexicons
- Sentence Alignment
  - Sentence Alignement Process
  - Cross-language Information Retrieval
  - Multilingual Linguistic Analysis
- Word Alignment
  - Single-Word Alignment
  - Compound-Word Alignment
  - Collocation Alignment
- Experimental Results
- Conclusion and Future Work

Construction of Translation Lexicons

- **Context and Objective**
  - ➢ Translation lexicons are vital in machine translation and cross-language information retrieval
  - ➢ The high cost of bilingual lexicon development and maintenance is a major entry barrier for adding new languages pairs

- **Approach for automatic construction**
  - ➔ Sentence alignment from parallel corpora
  - ➔ Word alignment from parallel corpora aligned sentence by sentence
  - ➔ Cleaning word alignment results

Outline

- Parallel Texts
- Construction of Translation Lexicons
- Sentence Alignment
  - ➢ Sentence Alignement Process
  - ➢ Cross-language Information Retrieval
  - ➢ Multilingual Linguistic Analysis
- Word Alignment
  - ➢ Single-Word Alignment
  - ➢ Compound-Word Alignment
  - ➢ Collocation Alignment
- Experimental Results
- Conclusion and Future Work

Sentence Alignment Approaches

- **Length-based approaches:** Short sentences will be translated as short sentences and long sentences as long sentences

- **Offset alignment by signal processing techniques:** These approaches attempt to align position offsets in the two parallel texts

  ➔ The goal is to induce an alignment by using cognates - words with similar forms and meanings across languages - at the level of character sequences

- **Lexical methods:** Use lexical information to align parts of sentences

Our Approach for Sentence Alignment

- **Objective**

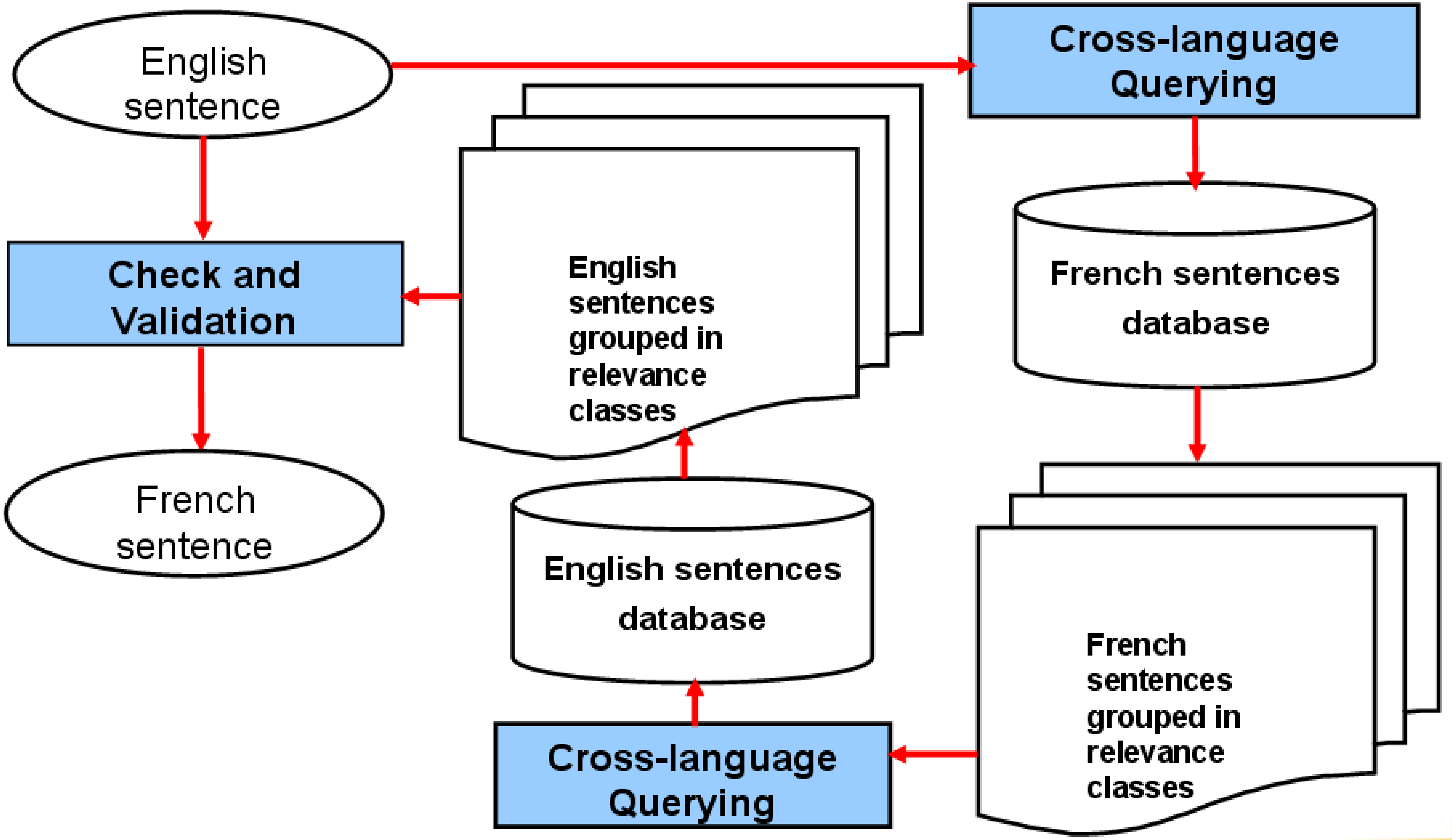
Combine different sources of information (bilingual lexicons, sentence length, sentence position, etc.) to improve performance of sentence alignment from parallel corpus

- **Principles**

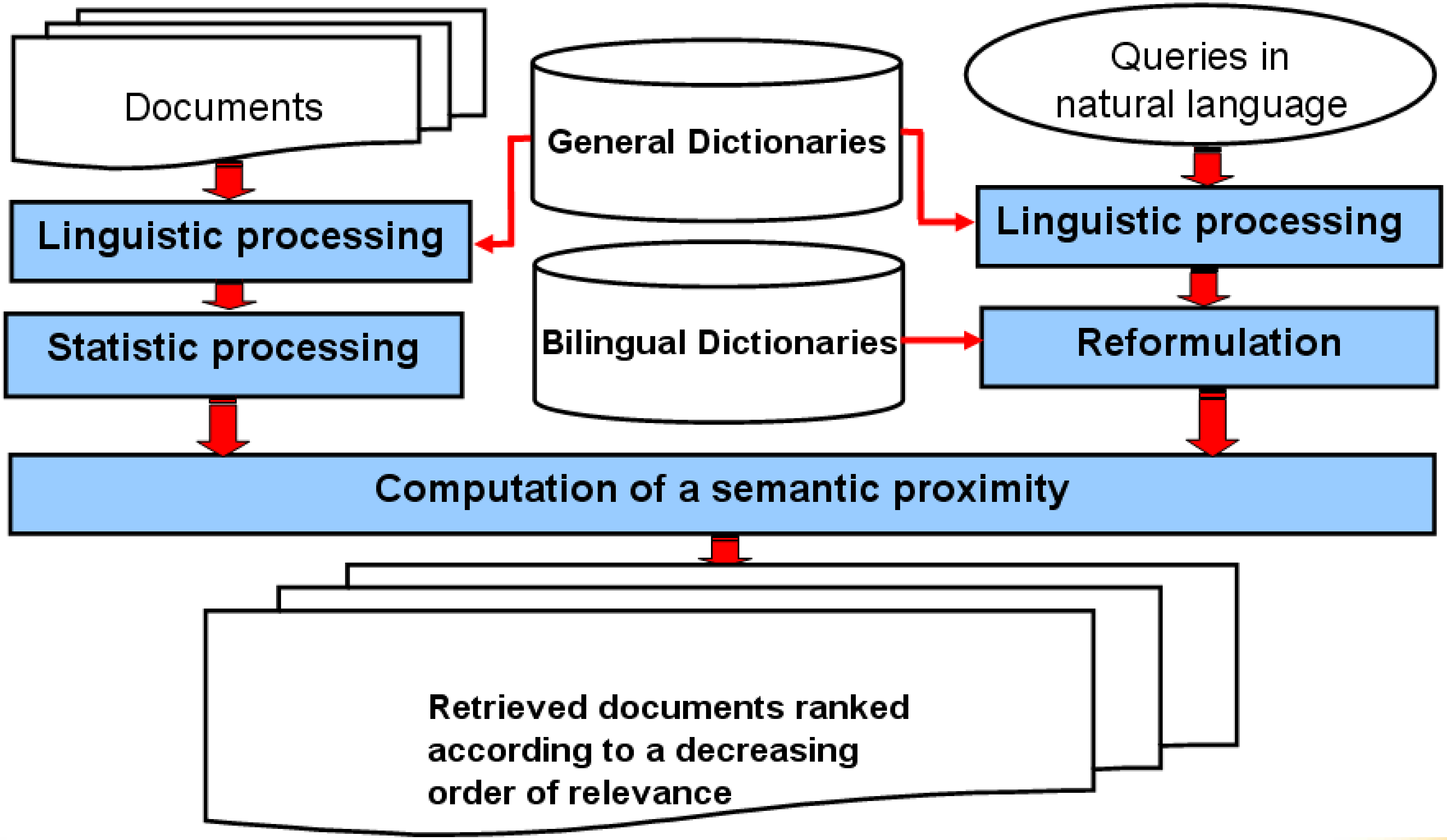
➔ Use of cross-language information retrieval techniques to estimate which sentence or sentences in source language correspond with which sentence or sentences in the target language

➔ Check and validate alignment by using criteria on common words of source and target sentences, sentence length and sentence position

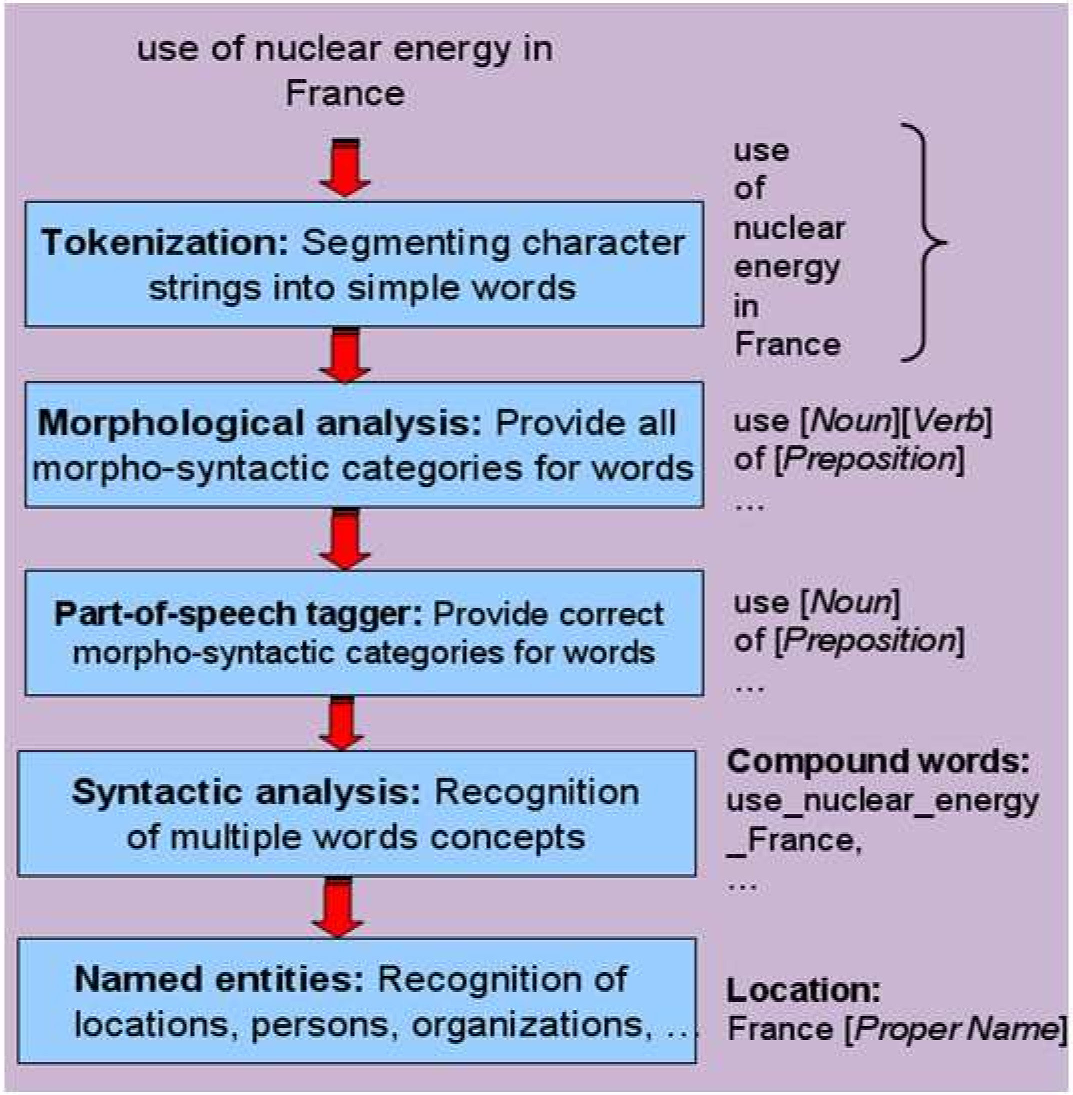
Sentence Alignment Process


English sentence
Cross-language Querying
Check and Validation
English sentences grouped in relevance classes
French sentences database
French sentence
English sentences database
French sentences grouped in relevance classes
Cross-language Querying

Cross-language Information Retrieval


Documents
General Dictionaries
Queries in natural language
Linguistic processing
Linguistic processing
Statistic processing
Bilingual Dictionaries
Reformulation
Computation of a semantic proximity
Retrieved documents ranked according to a decreasing order of relevance

Linguistic Processing



use of nuclear energy in France

Tokenization: Segmenting character strings into simple words

use
of
nuclear
energy
in
France

Morphological analysis: Provide all morpho-syntactic categories for words

use [*Noun*][*Verb*]
of [*Preposition*]
…

Part-of-speech tagger: Provide correct morpho-syntactic categories for words

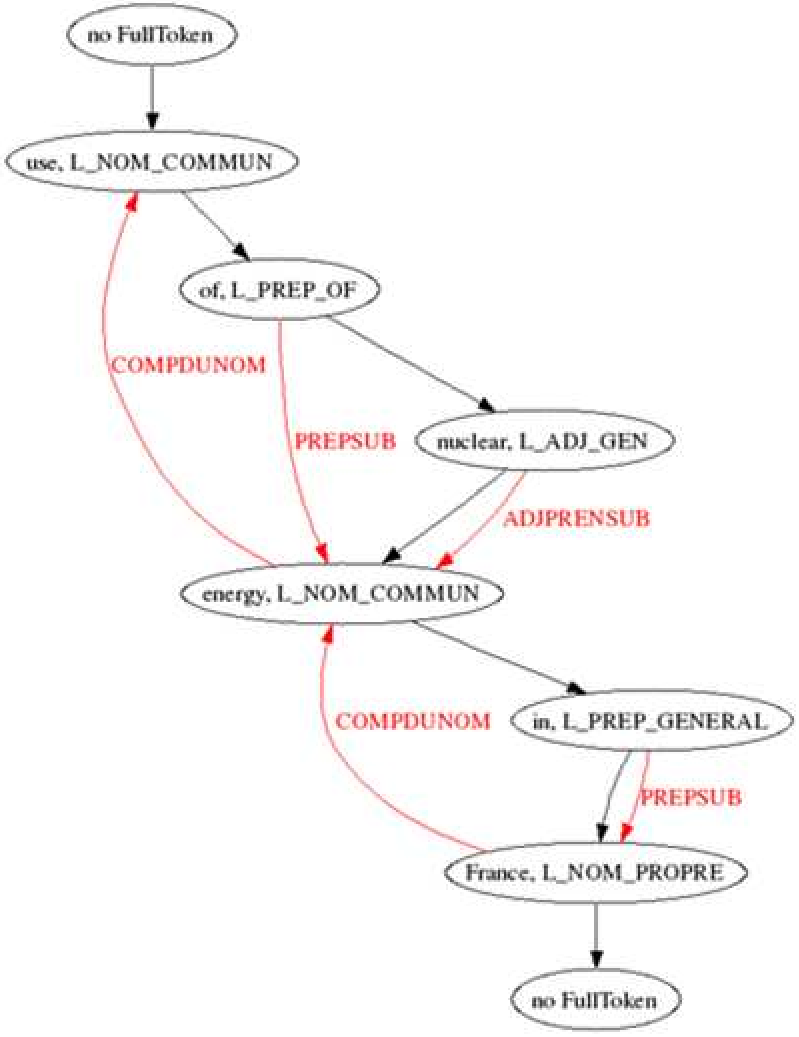
use [*Noun*]
of [*Preposition*]
…

Syntactic analysis: Recognition of multiple words concepts

Compound words:
use_nuclear_energy_France,
…

Named entities: Recognition of locations, persons, organizations, …

Location:
France [*Proper Name*]



Computation of Semantic Proximity between Documents and Queries

Statistical analysis

- Attribute a weight to each term t_i by using idf (inverse document frequency)

$$idf(t_i) = \log N/n_i$$

N: Number of documents in the database

n_i: Number of documents containing the term t_i

A term which occurs in many documents is not a good discriminator and should be given less weight that one which occurs in few documents

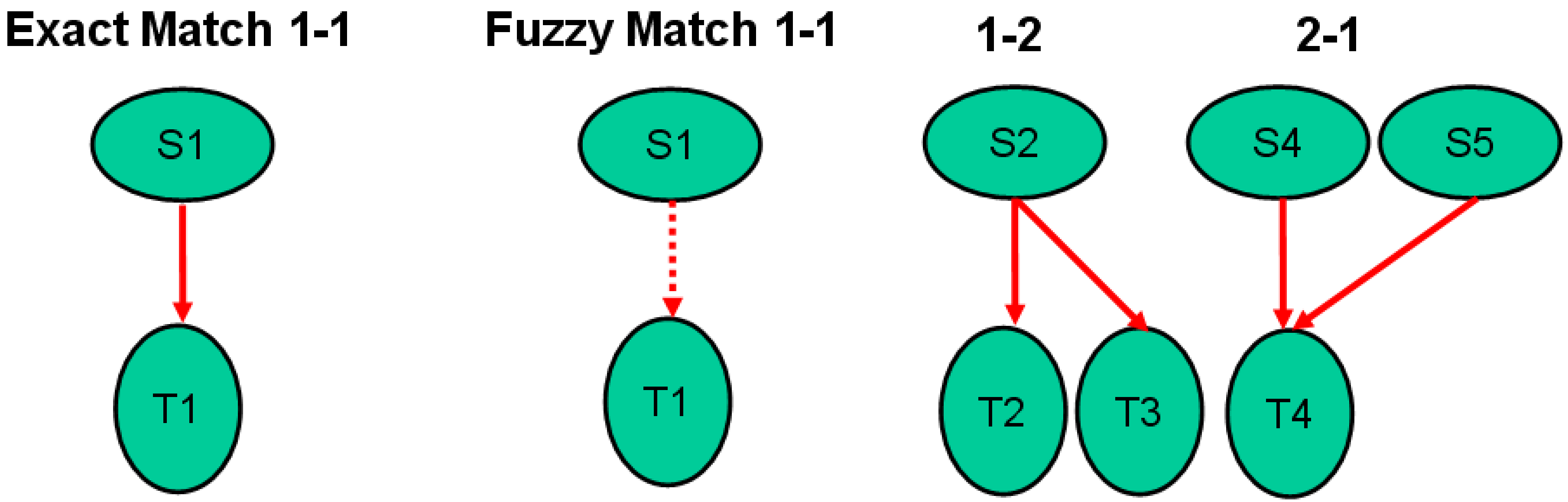
Semantic proximity computation

- Affect a relevance weight to the semantic intersection between query and documents

  Relevance weight = sum of the weights of terms present in documents

- Gather in one class documents with the same terms
- Sort classes according to their relevance

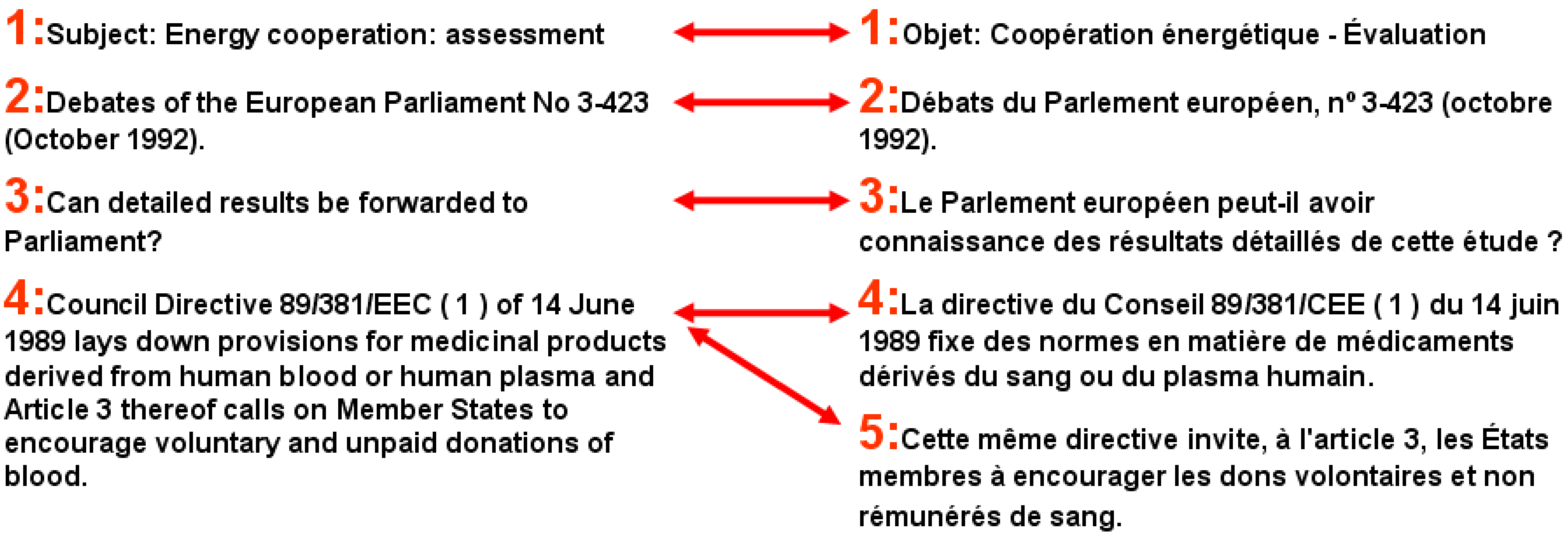
Sentence Alignment Types



Validation Criteria:

1. Position of the sentence in the corpus
2. Number of common words between the source sentence and the target sentence (semantic proximity)
3. Average ratio between the lengths of the source sentence and the target sentence (translation candidate)

Position of the sentence



1:Subject: Energy cooperation: assessment

2:Debates of the European Parliament No 3-423 (October 1992).

3:Can detailed results be forwarded to Parliament?

4:Council Directive 89/381/EEC (1) of 14 June 1989 lays down provisions for medicinal products derived from human blood or human plasma and Article 3 thereof calls on Member States to encourage voluntary and unpaid donations of blood.

1:Objet: Coopération énergétique - Évaluation

2:Débats du Parlement européen, nº 3-423 (octobre 1992).

3:Le Parlement européen peut-il avoir connaissance des résultats détaillés de cette étude ?

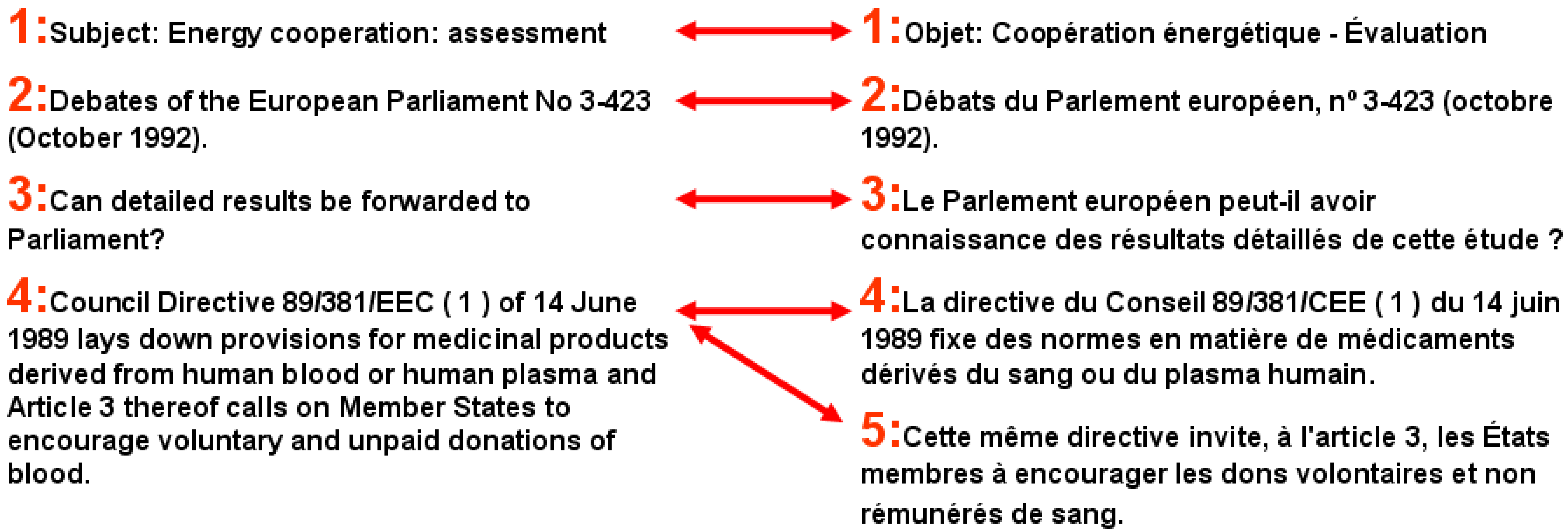
4:La directive du Conseil 89/381/CEE (1) du 14 juin 1989 fixe des normes en matière de médicaments dérivés du sang ou du plasma humain.

5:Cette même directive invite, à l'article 3, les États membres à encourager les dons volontaires et non rémunérés de sang.

- The position of the target sentence (translation candidate) must be located in a window of 10:

Last alignment – 5 <= **Position** <= Last alignment + 5

Length ratio average



1:Subject: Energy cooperation: assessment

2:Debates of the European Parliament No 3-423 (October 1992).

3:Can detailed results be forwarded to Parliament?

4:Council Directive 89/381/EEC (1) of 14 June 1989 lays down provisions for medicinal products derived from human blood or human plasma and Article 3 thereof calls on Member States to encourage voluntary and unpaid donations of blood.

1:Objet: Coopération énergétique - Évaluation

2:Débats du Parlement européen, nº 3-423 (octobre 1992).

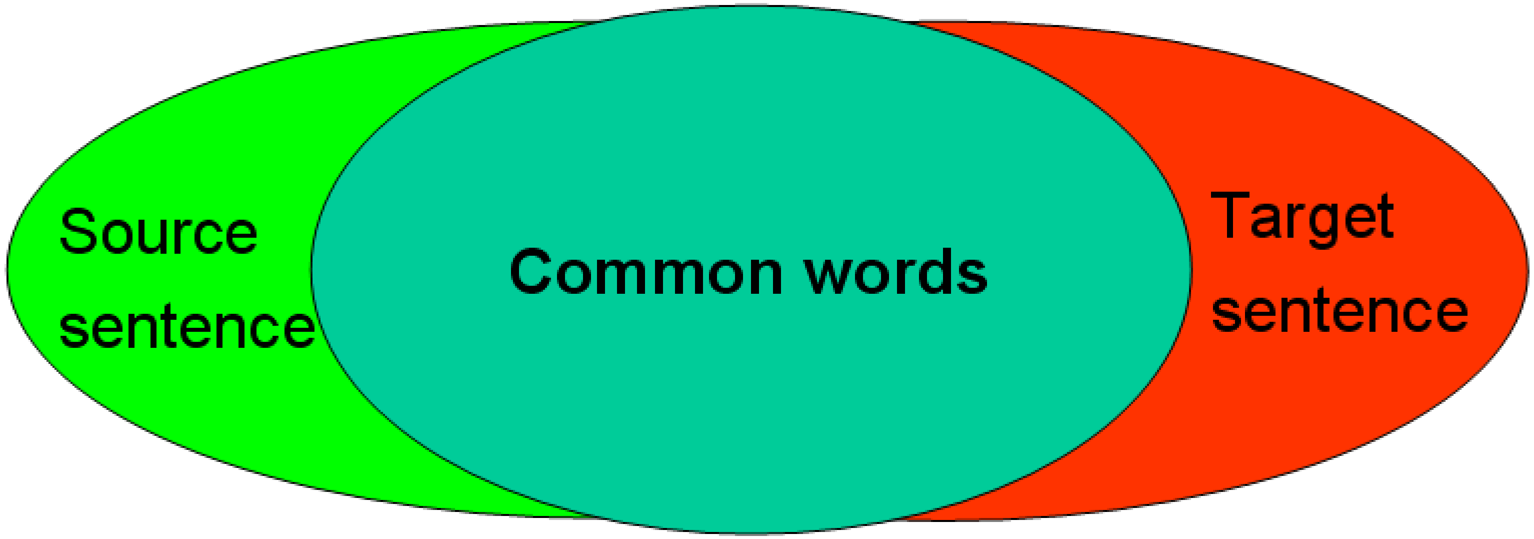
3:Le Parlement européen peut-il avoir connaissance des résultats détaillés de cette étude ?

4:La directive du Conseil 89/381/CEE (1) du 14 juin 1989 fixe des normes en matière de médicaments dérivés du sang ou du plasma humain.

5:Cette même directive invite, à l'article 3, les États membres à encourager les dons volontaires et non rémunérés de sang.

- μ: **The average ratio between the lengths of the source sentence and the target sentence:**
  - μ can be estimated by the ratio between the document length, since the majority of the sentences are 1 to 1
  - English/French: $0.95 <= \mu <= 1.25$

Number of common words



- The number of common words between the source sentence and the target sentence (semantic proximity) must be more than 50% of the number of words of the target sentence

Example: Exact Match 1-1 Alignment

- **The English sentence to align [75/1122]** "Social security funds in Greece are calling for independence with regard to the investment of capital"

Class number	Class query terms (Common words)	Class weight	Number of Retrieved sentences	Retrieved sentences
1	Greece, fund_security_social, independence, investment_capital	0.971079	1	[85/1127] L'indépendance, en ce qui concerne la gestion et l'investissement des fonds des caisses de sécurité sociale en Grèce, est une question d'ordre interne qui échappe à toute compétence communautaire
2	Social, Security, fund, Greece, fund_security_social, independence, investment	0.808404	1	[77/1127] Les caisses de sécurité sociale de Grèce revendiquent l'indépendance en matière d'investissements

Example: Exact Match 1-1 Alignment

- **The French sentence [77/1127]** “Les caisses de sécurité sociale de Grèce revendiquent l'indépendance en matière d'investissements” **is used as a query to the English database**

Class number	**Class query terms (Common words)**	**Class weight**	**Number of retrieved sentences**	**Retrieved sentences**
1	caisse_sécurité_social	0.789538	1	[75/1122] Social security funds in Greece are calling for independence with regard to the investment of capital
2	objet, social, caisse_sécurité, caisse_grec	0.380696	1	[74/1122] Subject: Greek social security funds

Example: Exact Match 1-1 Alignment

The English sentence "*Social security funds in Greece are calling for independence with regard to the investment of capital*"

is aligned to

the French sentence "*Les caisses de sécurité sociale de Grèce revendiquent l'indépendance en matière d'investissements*"

because:

- ➢ The position of the French sentence [77/1127] is included in a window of 10:
  (74 – 5) < 77 < (74 + 5) [Last alignment = 74]

- ➢ The number of common words between the English sentence and the French sentence is more than 50% of the number of words of the French sentence:
  6 > 8/2

- ➢ The length of the English sentence = μ * the length of the French sentence:
  100/96 = 1.04 (0.95 < μ = 1.04 < 1.25)

Outline

- Parallel Texts
- Construction of Translation Lexicons
- Sentence Alignment
  - ➢ Sentence Alignement Process
  - ➢ Cross-language Information Retrieval
  - ➢ Multilingual Linguistic Analysis
- Word Alignment
  - ➢ Single-Word Alignment
  - ➢ Compound-Word Alignment
  - ➢ Collocation Alignment
- Experimental Results
- Conclusion and Future Work

Word Alignment Approaches

- Statistical approaches based on IBM models

- Linguistic approaches for simple words and compound words alignment using bilingual lexicons and morpho-syntactic analysis on source and target sentences

- A combination of the two previous approaches

Single-Word Alignment Using Bilingual Dictionary

Approach: Look for the appropriate translation into the existing bilingual dictionary for each word of the source sentence

Example:

Source sentence : "The Commission considers that harmonization at Community level in this area is not necessary."

Target sentence: "La Commission considère qu'une harmonisation au niveau communautaire dans ce domaine n'est pas nécessaire. "

Source word	Position of the word in the sentence	Translation found in the dictionary	Target word	Position of the target word
commission	1		commission	1
consider	2	considérer	considérer	2
harmonization	3	harmonisation	harmonisation	3
level	4	niveau	niveau	4
community	5		communautaire	5
area	6	domaine	domaine	6
necessary	7		nécessaire	8

Single-Word Alignment Using Cognates

Approach: Search cognates (words having the same first four characters) among not assigned target words

Example:

Source sentence : "The Commission considers that harmonization at Community level in this area is not necessary."

Target sentence: "La Commission considère qu'une harmonisation au niveau communautaire dans ce domaine n'est pas nécessaire. "

Source word	Position of the word in the sentence	Translation found in the dictionary	Target word	Position of the target word
commission	1		commission	1
consider	2	considérer	considérer	2
harmonization	3	harmonisation	harmonisation	3
level	4	niveau	niveau	4
community	5		communautaire	5
area	6	domaine	domaine	6
necessary	7		nécessaire	8

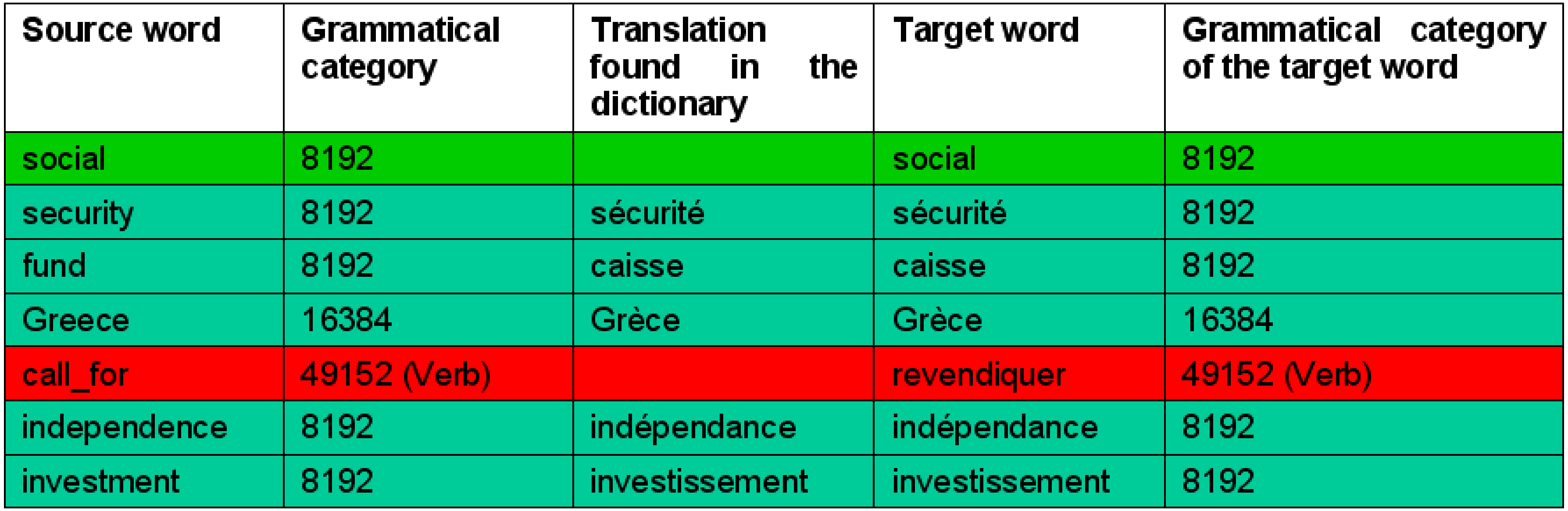
Single-Word Alignment Using Grammatical Tags

Approach: Search words having the same grammatical category

Example:

Source sentence : "Social security funds in Greece are calling for independence with regard to the investment of capital."

Target sentence: "Les caisses de sécurité sociale de Grèce revendiquent l'indépendance en matière d'investissements."



Source word	Grammatical category	Translation found in the dictionary	Target word	Grammatical category of the target word
social	8192		social	8192
security	8192	sécurité	sécurité	8192
fund	8192	caisse	caisse	8192
Greece	16384	Grèce	Grèce	16384
call_for	49152 (Verb)		revendiquer	49152 (Verb)
independence	8192	indépendance	indépendance	8192
investment	8192	investissement	investissement	8192

Compound-Word Alignment

Difficulties to align compound words and idiomatic expressions:

- A compound word is not automatically translated with a compound word.

  "Computer science" is translated as a single word "informatique"

- The translation of a compound word is not always obtained by translating its components separately.

  "fuel saving" is not translated as "économie de carburant" but "économie d'énergie"

- A same compound word can have different forms due to the morphological, syntactic and semantic changes. These changes must be taken into account in the alignment process.

  "water resources management" and "management of water resources" have the same translation "gestion des resources en eau"

Compound-Word Alignment

Alignment of compound words which are translated word to word:

- Apply a syntactic analysis on the source and target sentences in order to extract dependency relations between words and to recognize compound words structures
- Apply reformulation rules on compound words structures in order to establish correspondences between the compound words of the source sentence and the compound words of the target sentence

Example of a reformulation rule between English and French:

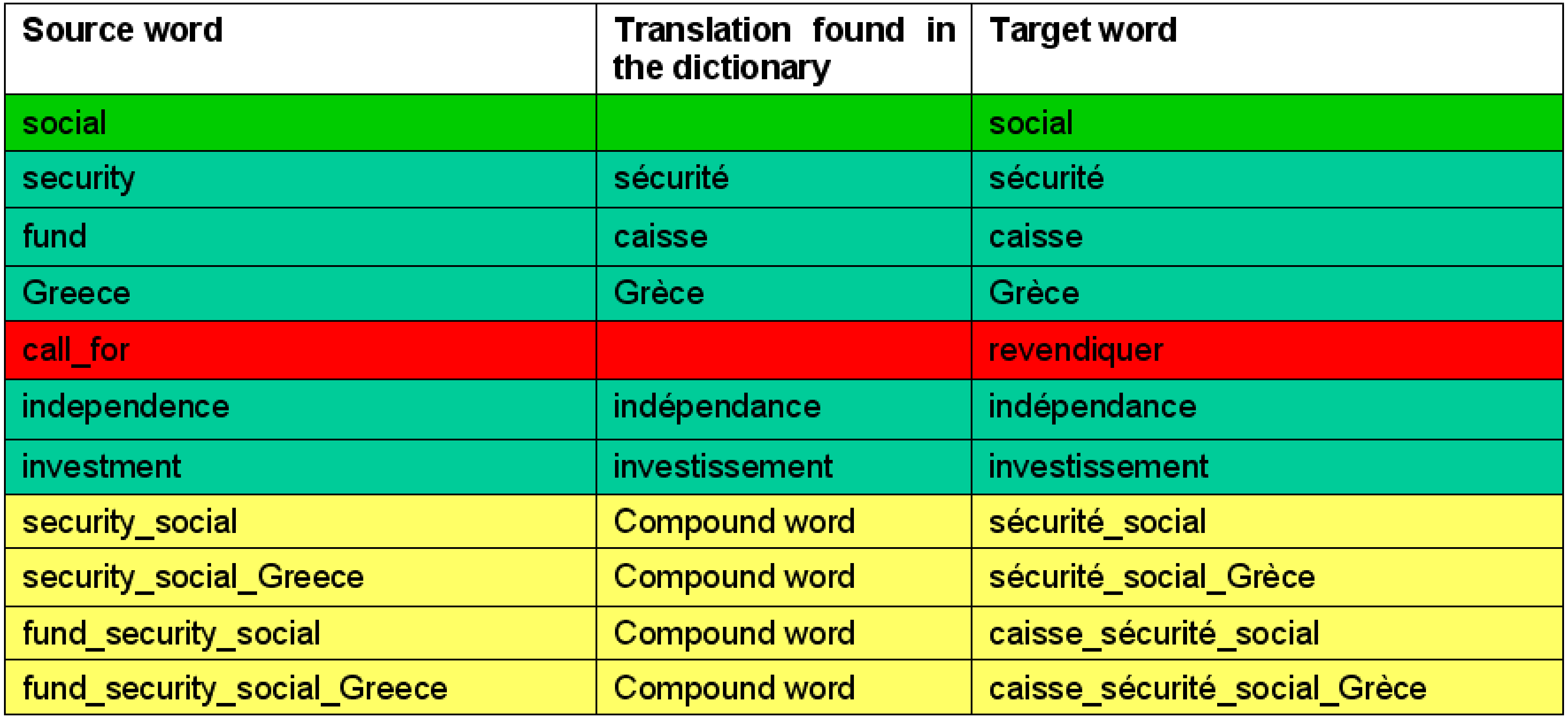
- *Translation(**A.B**) = Translation(**B**).Translation(**A**)*

*Translation(**green.mountain**) = Translation(**mountain**).Translation(**green**)*

= ***montagne.verte***

Compound-Word Alignment

Approach: Establish automatically correspondences between compound words of the source and target sentences



Source word	Translation found in the dictionary	Target word
social		social
security	sécurité	sécurité
fund	caisse	caisse
Greece	Grèce	Grèce
call_for		revendiquer
independence	indépendance	indépendance
investment	investissement	investissement
security_social	Compound word	sécurité_social
security_social_Greece	Compound word	sécurité_social_Grèce
fund_security_social	Compound word	caisse_sécurité_social
fund_security_social_Greece	Compound word	caisse_sécurité_social_Grèce

Collocation Alignment

- Works on comparable corpora with paragraph (segment) level alignment
- Rests on a monolingual MWE extraction in both languages
- Useful on noisy data such as corpora crawled from the Web

Algorithm:

- Simple premise :
  - An MWE is translated by an MWE
  - Corresponding MWEs should appear in the same place in the same segments
- Monolingual MWE extraction
- Translations should have the same frequency. We can define a simple threshold,

Frequency distance: $$Fd = \frac{|f(s) - f(t)|}{\max(f(s), f(t))}$$

Collocation Alignment

Algorithm:

- On top of appearing a similar number of time, translations should appear in the same segments. A second threshold is applied

Co-occurrence distance: $$Cd = \frac{\sqrt{\Sigma(X_i - Y_i)^2}}{N}$$

- Heuristics: usually corresponding MWEs have a similar length, if it is too different, the alignment is discarded

Outline

- Parallel Texts
- Construction of Translation Lexicons
- Sentence Alignment
  - Sentence Alignement Process
  - Cross-language Information Retrieval
  - Multilingual Linguistic Analysis
- Word Alignment
  - Single-Word Alignment
  - Compound-Word Alignment
  - Collocation Alignment
- **Experimental Results**
- Conclusion and Future Work

Experimental Results: English-French Sentence Alignment

Data:

- Part of the ARCADE II corpus (Official Journal of the European Community)
- 1 103 English sentences aligned to their French counterparts (1 122 English sentences, 1 127 French sentences)

Alignment process:

1. **Exact Match 1-1 Alignment** (Use of the three alignment criteria: Number of common words between the source and target sentences; Position of the target sentence; Ratio of lengths of the source and target sentences)
2. **1-2 Alignment** (Use of the first two alignment criteria)
3. **2-1 Alignment** (Use of the first two alignment criteria)
4. **Fuzzy Match 1-1 Alignment** (No use of alignment criteria: Alignments which are partially correct)

Experimental Results: English-French Sentence Alignment

$S = \{s_1, s_2, \ldots, s_n\}$: Source text

$T = \{t_1, t_2, \ldots, t_m\}$: Target text

➔ An alignment **A** is a sub-set of the Cartesian product S x T

➔ The comparison of an alignment **A** with a reference alignment $\mathbf{A}_r$ is achieved by computing the area of the intersection of **A** and $\mathbf{A}_r$ in terms of number of characters

$$\text{Precision} = \frac{\text{Area}(A \cap Ar)}{\text{Area}(A)}$$

$$\text{Recall} = \frac{\text{Area}(A \cap Ar)}{\text{Area}(Ar)}$$

$$\text{F - measure} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precison} + \text{Recall})}$$

	Precision	**Recall**	**F-measure**
Sentence aligner	0.99	0.99	0.99

Experimental Results: Single and Compound Word Alignment

$$\text{Precision} = \frac{|A \cap A_r|}{|A|} \qquad \text{Recall} = \frac{|A \cap A_r|}{|A_r|} \qquad \text{F - measure} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precison} + \text{Recall})}$$

A: set of alignments provided by the word aligner

Ar: corresponds to the set of the correct alignments

	Precision	**Recall**	**F-measure**
Single-word aligner	0.90	0.81	0.85
Compound-word aligner	0.84	0.55	0.66

➔ 54% of words are aligned with the bilingual lexicon, 8% are aligned with cognates detection and 26% are aligned by using grammatical tags

➔ 34% of the words of the source sentence and their translations are added to the bilingual lexicon

Experimental Results: French-English Collocation Alignment

Data:

- Hansard corpus
- 302 000 French sentences aligned to their English counterparts

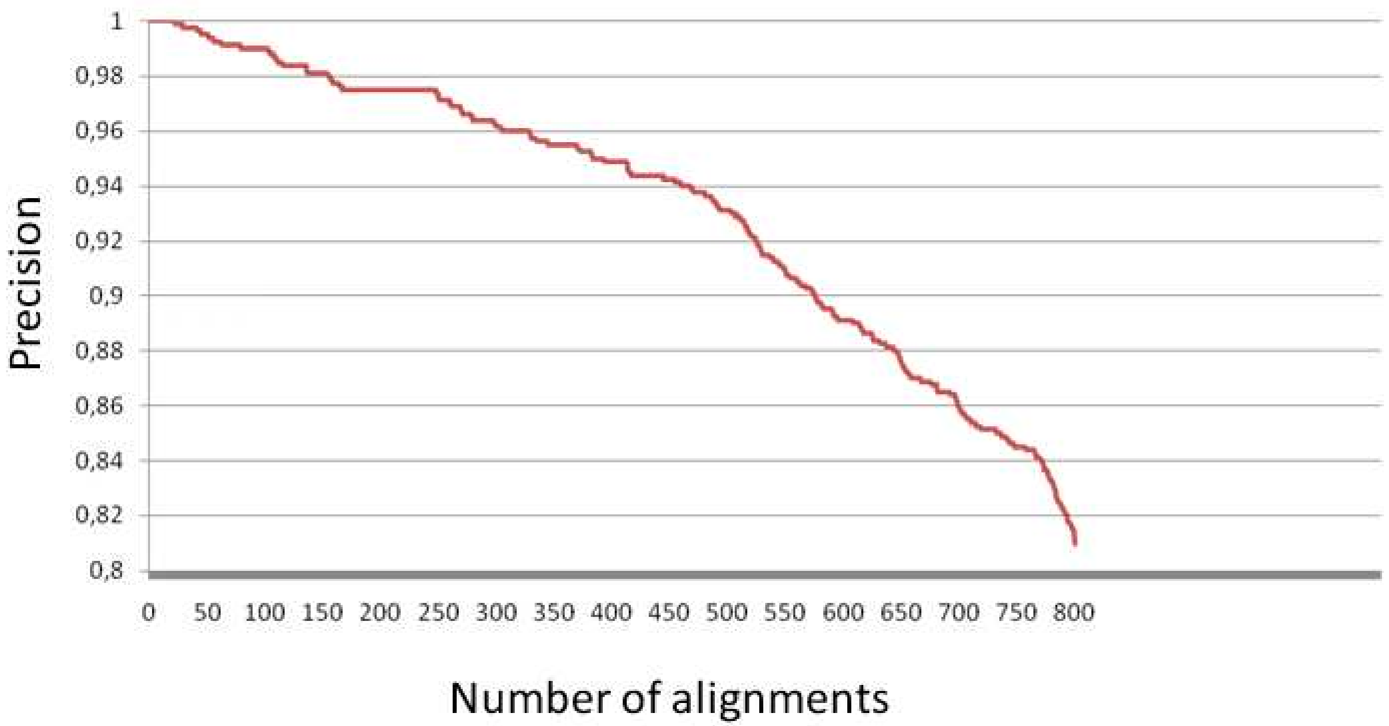
	Precision	Recall	F-measure
Collocation aligner	0.81	0.38	0.52

Examples:

- opposition officielle ➔ official opposition
- taux de intérêt ➔ interest rates
- vache à lait ➔ cash cow

Experimental Results: French-English Collocation Alignment

How does the term frequency affect the precision?



Outline

- Parallel Texts
- Construction of Translation Lexicons
- Sentence Alignment
  - Sentence Alignement Process
  - Cross-language Information Retrieval
  - Multilingual Linguistic Analysis
- Word Alignment
  - Single-Word Alignment
  - Compound-Word Alignment
  - Collocation Alignment
- Experimental Results
- Conclusion and Future Work

Conclusion and Future Work

- **Conclusion**
  - → High precision and recall for single-word alignment
  - → State of the art precision for compound-word and collocation alignment
  - → Deep linguistic analysis improves performance of sentence alignment and word alignment
  - → Still not enough for fully automated lexicon creation
- **Future Work**
  - → Improve the recall of the compound-word and collocation alignment by adding linguistic knowledge to the process
  - → Develop an ergonomic user interface for linguists