Outline

Agenda

Paraphrases

Corpus

Previous Work

Methodology

Logos Model

CLUE

Paraphrasary

eSPERTo

Summary

Future Work

A Multilingual Paraphrasary of Multiwords

Anabela Barreiro
anabela.barreiro@inesc-id.pt

Cristina Mota
cristina.mota@inesc-id.pt

Multi3Generation (CA18231)
Multilingual, Multimodal and Multitask Language Generation

Tampere – Finland, 15th June, 2023
Co-located with The 24th Annual Conference of The European Association for Machine Translation

Funded by the Horizon 2020 Framework Programme of the European Union

Agenda

Outline

Agenda

Paraphrases

Corpus

Previous Work

Methodology

Logos Model

CLUE

Paraphrasary

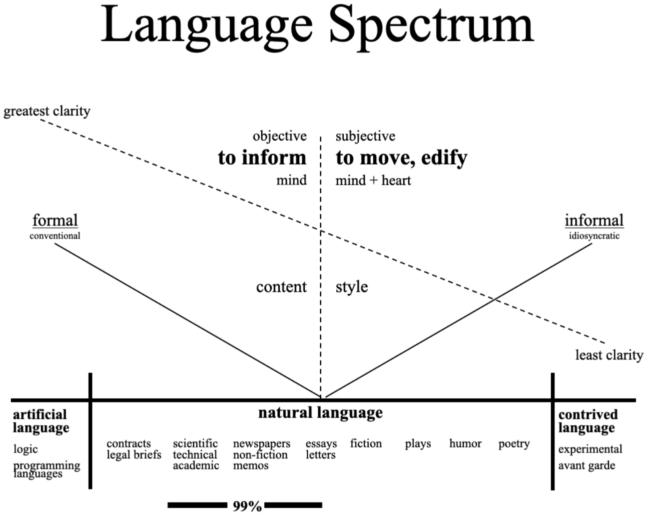
eSPERTo

Summary

Future Work

- Paraphrases
- Corpus
- Previous Work
- Methodology
- Alignments
- The Logos Model
- CLUE – **C**ross-**l**ingual **U**nit **E**licitation
- Paraphrasary
- eSPERTo – **S**ystem for **P**araphrasing in **E**diting and **R**evision of **T**exts
- Summary
- Future Work

Paraphrases



Paraphrases

- Paraphrasing means rephrasing or rewording
- It is a technique that consists of changing the words and structure (the form, syntax) of a text, yet preserving its meaning (the semantics)
- Paraphrases are often sought by humans when searching for different ways to articulate the same meaning or idea even in larger contexts
- Paraphrase generation is crucial in natural language processing (NLP)
- Collecting a dataset of paraphrases to be used in NLP and natural language generation (NLG) tasks including machine translation (MT)
- Different levels and categories of paraphrases
  - Barreiro (2009), Chapter 2, pp. 21-32
- Approximate paraphrase (‘approximate meaning’) is a constant challenge
  - Barzilay and McKeown (2001)

Paraphrases

- Our understanding and definition of paraphrase covers a **broad spectrum of semantico-syntactic phenomena** which correspond to **exact or approximate equivalences in meaning** among them and can range from short multiwords to sentences, or spread to contexts larger than the sentence level
- Here, we restrict the analysis of paraphrases to multiwords – level smaller than the sentence

Paraphrases

- what makes two sentences a paraphrase of each other?
- is it enough to change a word in a sentence to generate a paraphrase?
- to what extent can a sentence conveying the same idea be considered a paraphrase?
- How much semantic deviation is accepted to still consider two sentences or two expressions paraphrases of each other?
- Which paraphrases are useful and productive and which ones cause more harm than good, computationally-wise?

Paraphrases and/in Translation

- Translation is paraphrasing in different languages
- In the context of **translation**, paraphrases represent different but semantically similar ways of translating the same expression or sentence
- Paraphrases are vital to deploy semantic knowledge to guarantee high fidelity translation
- An important common issue in human translation and MT is to define equivalence and establish paraphrasing capabilities
- Quality machine translation (MT) cannot be achieved without comparable quality paraphrase knowledge
- One of the first tasks involved in the construction of a paraphrasing or MT system should be to collect pairs of alignments which correspond to semantically identical or similar units of meaning expressed with different vocabulary and/or syntactic structure

Corpus

- **Common test version of the European Parliament Proceedings** taken from Q4/2000 portion of the data, 2000-10 to 2000-12 (Koehn, 2005)
- The bilingual texts are available at the European Parliament Proceedings Parallel Corpus website `http://www.statmt.org/europarl/archives.html#v1`
- The reference sub-corpus is **aligned at the sentence level**, ranging from sentence number 101 to sentence number 500
- Our work represents an extension of the work on **multilingual alignments** (Graça et al., 2008)
- We manually annotated translation alignments for **400 sentences in 6 sets of the multilingual test corpus**, representing 2,400 aligned sentences
- In the corpus not all translations are optimal and often translational equivalents are approximate rather than exact

Previous Work

- "n-grams in search of theories" (Maia et al., 2008) – claimed the need to create linguistically robust alignment tools for research based on a supporting theoretical and practical framework
- CLUE-Aligner (Barreiro et al., 2016) appeared as a response to the demand for the alignment of contiguous and non-contiguous multiwords, i.e., units with insertions, such as support verb constructions (Barreiro and Batista, 2016)
  - EN – to draw a distinction between | PT – estabelecer uma diferença entre
  - EN – to bring [INSERTION] to a conclusion | PT – concluir
- Lately, we have been addressing multiwords **contextual nuances**, such as the prepositional verb *break into* in the EN-PT alignment pairs, a challenge that was addressed successfully in the Logos Model
  - EN – break into NPL | PT – assaltar NPL (beak into a house — assaltar uma casa)
  - EN – break into a laugh | PT – desatar a rir
  - EN – break into a run | PT – pôr-se em fuga, pôr-se a andar, pôr-se a milhas
  - EN – break into pieces | PT – quebrar em bocados, estrilhaçar

Methodology

- Identification of 4 classes of challenges to the alignment of multiwords
  - lexical and semantico-syntactic
  - morphological
  - morpho-syntactic
  - semantico-discursive
- Focus on the **lexical and semantico-syntactic phenomena** that MT systems, in general, do not translate well, namely the alignment of multilingual/crosslingual multiwords
- From the alignment task, resulted a paraphrase collection to be used in distinct types of NLP applications including MT
- Analysis of the collection and creation of a novel linguistic computational object/concept – **Paraphrasary**

Methodology

- In order to achieve a provisional first-round of results, a polyglot linguist, with knowledge of the 4 languages covered in this study, annotated manually the total of 2,400 sentence alignments (400 x 6 language pairs) and built the CLUE-Guidelines based on linguistic knowledge as processed in the Logos Model, paying special attention to multiwords and other translation units
- From the dataset of 400 sentences of the corpus, for the EN-PT language pair, a total of 3,700 multiword alignments were collected. They all represent candidates for entries in our Paraphrasary

Alignments of Multiwords

- CLUE-Alignment Guidelines – based on the fundamental principles of the Logos Machine Translation Model (henceforth, the **Logos Model**) (Scott, 2003) (Barreiro et al., 2011), which relies on **deep semantico-syntactic analysis to generate translation of multiwords**
  - EN – give in without struggle | PT – ceder sem resistência
  - EN – NHum/PRO be settling down to PRO new job | PT – NHum/PRO ir-se habituando ao novo emprego
  - EN – arrive first/second/last | PT – chegar em primeiro/segundo/último lugar
- Quality texts and quality alignments based on the "Semtab" function of the Logos Model were key ingredients to build an efficient **multilingual paraphrasary**, which represents a step forward into meaningful quality translation, and a valuable resource for NLG and MT

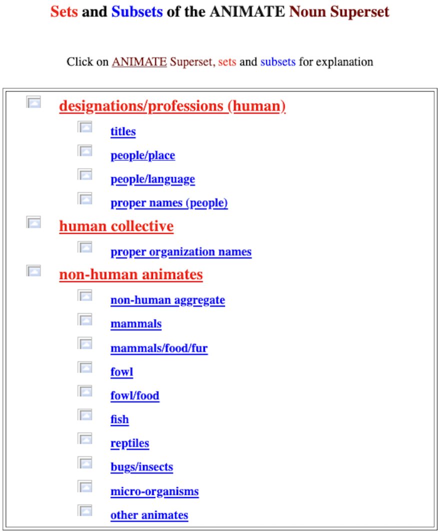
The Logos Model Approach

- The Logos Model has been described with a great degree of detail in (Scott, 2003), (Scott and Barreiro, 2009), and (Scott, 2018), among others.
- We highlight in this paper only the SAL language and the SemTab function for the sake of illustrating how relevant they are to our approach on the processing and generation of multiwords and the establishment of bilingual and multilingual paraphrasaries
  - SAL
  - SemTab
- An area in which fewer research efforts have been made
- It is crucial for higher quality paraphrasing translation
- It is used in a large range of NLP applications

The Logos Model Approach – SAL

- Natural language is represented as a refined Semantico-Syntactic Abstraction Language (SAL), also designated as a hierarchical ontology
- Categories for all parts of speech
- When processing the sentence, words strings are converted into SAL patterns
- SAL has 4 levels of abstraction: 1 syntactic level (word class), and 3 levels
  - superset
  - set
  - subset
- It is possible to apply the same techniques to the data, which in the Logos Model are not literal words, but SAL entities or SAL patterns
- This is the reason why it makes sense to train machine learning (ML) systems to learn new SAL patterns based on alignments, instead of on the conventionally-used MT patterns

The Logos Model Approach – SAL



Sets and Subsets of the ANIMATE Noun Superset

Click on ANIMATE Superset, sets and subsets for explanation

- designations/professions (human)
  - titles
  - people/place
  - people/language
  - proper names (people)
- human collective
  - proper organization names
- non-human animates
  - non-human aggregate
  - mammals
  - mammals/food/fur
  - fowl
  - fowl/food
  - fish
  - reptiles
  - bugs/insects
  - micro-organisms
  - other animates

Figure: SAL Superset Animate-type Nouns

The Logos Model Approach – SemTab

Outline

Agenda

Paraphrases

Corpus

Previous Work

Methodology

Logos Model

CLUE

Paraphrasary

eSPERTo

Summary

Future Work

- Multiwords are represented as rules in a separate database, the Semantic Table or "SemTab" (Orliac and Dillinger, 2003)
- The methodological framework for the alignment task relies on the use of multiwords as representation objects of alignment
- The meaning is derived from the semantic processing in the SemTab function, where multiwords can be processed and translation fidelity can be improved
- SemTab allows to distinguish between the multiword
  - EN – be acquainted with N(AN-Hum)/PRO | PT – conhecer N/PRO pessoalmente
    the translation of the verb depends on the type of noun (human-type) and the multiword
  - EN – be acquainted with N-Abs | PT – estar ao corrente de N-Abs
    the noun N is abstract (Abs) of the type "Information", e.g., a piece of news, a gossip, situation, etc..

The Logos Model Approach – SemTab

- EN – he was driving the car at full speed, the noun car can be replaced by any type of concrete, vehicle: drive N(CO-Vehic) at full speed
  PT – guiar/conduzir N(CO-Vehic) a toda a velocidade
  (Power of generalisation)
- SemTab rules deploy SAL patterns or entities, such as the aforementioned N(AN-Hum), N(Abs-info-type), or N(COVehic)

The Logos Model Approach – SemTab

- In the example, the English prepositional verb *to deal with* is translated in the Romance languages as - *dedicarse a* (engage in) in Spanish, - the reflexive *s'attacher à* (focus on/stick to) in French, - and *centrar-se em* (concentrate/center/focus on) in Portuguese
  - EN - our Asian partners prefer to deal with questions which unite us
  - ES - nuestros socios asiáticos prefieren dedicarse a las questiones que nos unen
  - FR - nos partenaires asiatiques préfèrent s'attacher à ([a+a]) ce qui nous unit
  - PT - os nossos parceiros asiáticos preferem centrar-se unicamente nas ([em+as]) questões comuns
- SemTab rules deploy SAL patterns or entities, such as the aforementioned N(AN-Hum), N(Abs-info-type), or N(COVehic)
- Application of a SemTab contextual rule, such as the one below, which is a deep structure pattern that matches on/applies to a great variety of surface structures
  - EN – DEAL(VI) WITH N(questions) | PT – S'OCCUPER DE N

CLUE – CrossLingual Unit Elicitation

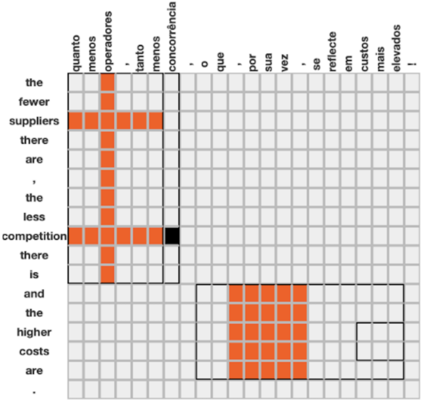
Under the umbrella of CLUE, we developed:

- Alignment guidelines (revision and refinement)
- An alignment tool: CLUE-Aligner
- Gold collection: Gold-CLUE (revision and refinement)

CLUE-Aligner

- Alignment tool developed to annotate paraphrasing or translation units representing multiwords found in monolingual or bilingual pairs of parallel sentences (Barreiro et al., 2016)
- Based on Linear-B (Callison-Burch and Bannard, 2004), extended in order to allow the alignment of contiguous and non-contiguous multiwords, addressing the long-distance dependency that characterizes the majority of semantico-syntactic patterns
- Allows loading of previously generated alignments (segments) for the corpora parallel sentences
- During the annotation task, the annotator manually corrects any inaccurate alignments (either gathered manually or automatically), and defines the new alignments for multiwords, which represent translation (or paraphrasing) units

CLUE-Aligner



	quanto	menos	operadores	,	tanto	menos	concorrência	,	o	que	,	por	sua	vez	,	se	reflecte	em	custos	mais	elevados	!
the			🟧																			
fewer			🟧																			
suppliers	🟧	🟧	🟧	🟧	🟧	🟧																
there			🟧																			
are			🟧																			
,			🟧																			
the			🟧																			
less			🟧																			
competition	🟧	🟧	🟧	🟧	🟧	🟧	■															
there			🟧																			
is			🟧																			
and											🟧	🟧	🟧	🟧	🟧							
the											🟧	🟧	🟧	🟧	🟧							
higher											🟧	🟧	🟧	🟧	🟧							
costs											🟧	🟧	🟧	🟧	🟧							
are											🟧	🟧	🟧	🟧	🟧							
.																						

the fewer [] there are , the less competition there is | quanto menos [] , tanto menos concorrência (S)

suppliers | operadores (S)

the fewer [] there are , the less [] there is | quanto menos [] , tanto menos (S)

competition | concorrência (S)

and the higher costs are | o que [] se reflecte em custos mais elevados (S)

higher costs | custos mais elevados (S)

CLUE-Aligner

Outline

Agenda

Paraphrases

Corpus

Previous Work

Methodology

Logos Model

CLUE

Paraphrasary

eSPERTo

Summary

Future Work

- Alignment of the noncontiguous comparison/metonymy
  - EN – the fewer [] there are, [] the less [] there is | PT – quanto menos [], tanto menos
  - EN – the higher [] are | PT – o que se reflecte em [] mais elevados
- insertions are excluded and aligned independently
  - EN suppliers | PT – operadores
  - EN competition | concorrência
  - EN costs | PT custos
- The linguistic annotator can immediately see the list of alignments in text format and correct any error that might have been done in the alignment task

Gold-CLUE

- Gold collection made of aligned multiwords resultant from our alignment task
- Contemplates a set of linguistic phenomena that can be classified into 4 main classes
  - **Lexical and semantico-syntactic** challenges include multiwords, such as support verb constructions, compound/modal verbs, and prepositional predicates
  - **Morphological** challenges include contracted forms, lexical versus non-lexical realization, that is, lexical items that are present in one language but not the other, such as determiners (articles and zero/missing articles), and pro-drop phenomena including subject pronoun dropping, and empty relative pronouns
  - **Morpho-syntactic** challenges include free noun adjuncts (noun-noun compounds)
  - **Semantico-discursive** challenges include emphatic linguistic constructions, such as pleonasm and tautology, repetition, and focus constructions

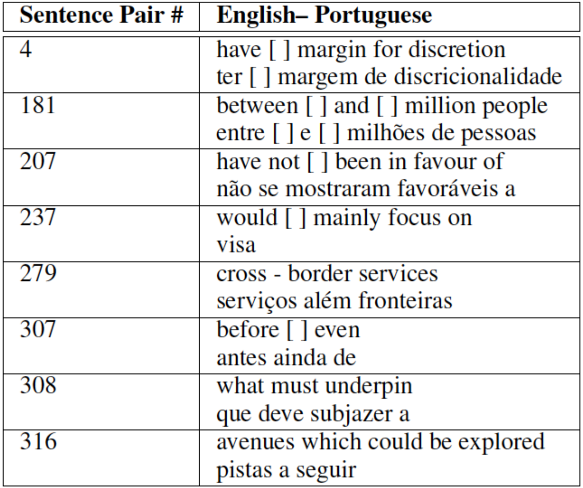
Outline

Agenda
Paraphrases
Corpus
Previous Work
Methodology
Logos Model
CLUE
Paraphrasary
eSPERTo
Summary
Future Work

Paraphrasary

- The objects of alignment provide a multilingual complex dictionary-type function, which we call **Paraphrasary**
- A paraphrasary is to semantico-syntactic units' equivalences as a standard dictionary is to synonyms
- It is a database of multiword entries listed alphabetically validated by a linguist after these multiwords have been aligned during the alignment task.
- Paraphrases generated via a paraphrasary clarify, simplify and add precision to text, and to its translation
- Our research on paraphrasing applications shows that both monolingual and multilingual paraphrase generation require the development of paraphrasaries
- Paraphrasary is a new concept of organizing linguistic data in a repository (or several repositories), which can grow into a large body of paraphrastic knowledge
- Indispensable in paraphrasing and (machine) translation

Paraphrasary Candidates



Sentence Pair #	English– Portuguese
4	have [ ] margin for discretion ter [ ] margem de discricionalidade
181	between [ ] and [ ] million people entre [ ] e [ ] milhões de pessoas
207	have not [ ] been in favour of não se mostraram favoráveis a
237	would [ ] mainly focus on visa
279	cross - border services serviços além fronteiras
307	before [ ] even antes ainda de
308	what must underpin que deve subjazer a
316	avenues which could be explored pistas a seguir

Table 1: EN-PT Alignments

Paraphrasary Candidates

- Alignment 181 in the Table above
  - EN – between [ ] and [ ] million people | PT – entre [ ] e [ ] milhões de pessoas

  can enter the Paraphrasary via a SemTab-type rule that allows to generate a large number of instances.
- an alignment can become much broader by using some constraints, such as [Num], a numeric expression
  - EN – between [NUM] and [NUM] N | PT – entre [NUM] e [NUM] de N
- Via the power of generalization that SAL categories allow, an alignment pair gathered from the corpus can be used in the generation of thousands of multilingual paraphrases
- The development of paraphrasaries is the **kick-start** of a paraphrasing tool

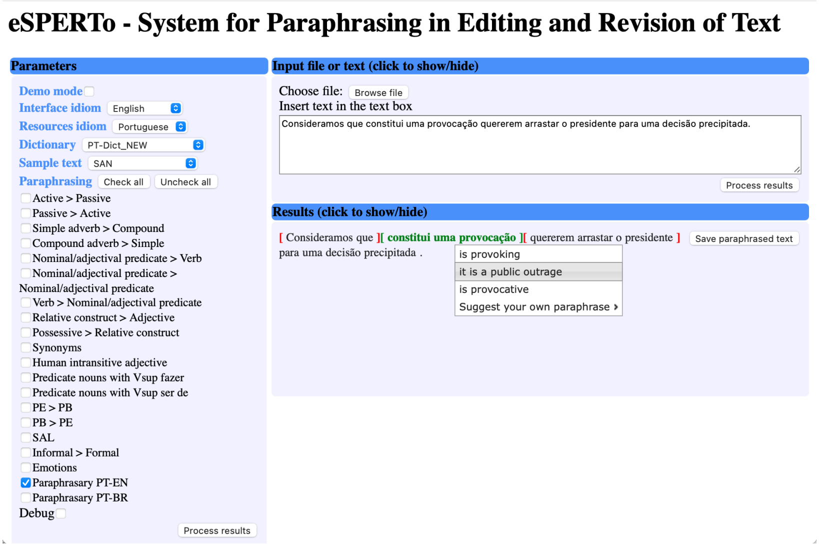
Multilingual Paraphrasaries

- Ongoing work carried out in compliance with the CLUE-Alignments, a set of linguistically informed multilingual alignments, comprising several categories of multiwords
  - The **CLUE-Alignments** set has all possible combinations between English, French, Portuguese, and Spanish parallel texts of the common test set of the Europarl corpus
- The gold collection of the annotated CLUE-Alignments is **Gold-CLUE**
- The **CLUE-Aligner** tool was developed to facilitate the alignment of the meaning and translation units in the bitexts, including the alignment of non-contiguous multiwords
- Our approach benefits from the **Logos Model** for machine translation, namely the semantico-syntactic abstraction language SAL and the semantic table function **SemTab**
- How the collected paraphrases are used in the paraphrase generation tool **eSPERTo**, developed for Portuguese, as part of a **larger multilingual generation project involving paraphrasing and translation**

eSPERTo Paraphrase Generation System

- **S**ystem for **P**araphrasing in **E**diting and **R**evision of **T**ext
- Online platform that allows rewriting different kinds of expressions using the NooJ linguistic engine (Silberztein, 2015; Silberztein, 2003)
- The system can be tested at: `https://esperto.hlt.inesc-id.pt/esperto/esperto/demo.pl`
- (Barreiro et al., 2022) present an overview of the system and lexicon-grammar resources that allow for the easy paraphrasing of constructions involving human intransitive adjectives, and also predicate nouns with support verbs *fazer* (do) and *ser de* (be of)

eSPERTo Paraphrase Generation System



eSPERTo - System for Paraphrasing in Editing and Revision of Text

Parameters

Demo mode □

Interface idiom: English

Resources idiom: Portuguese

Dictionary: PT-Dict_NEW

Sample text: SAN

Paraphrasing

- □ Active > Passive
- □ Passive > Active
- □ Simple adverb > Compound
- □ Compound adverb > Simple
- □ Nominal/adjectival predicate > Verb
- □ Nominal/adjectival predicate > Nominal/adjectival predicate
- □ Verb > Nominal/adjectival predicate
- □ Relative construct > Adjective
- □ Possessive > Relative construct
- □ Synonyms
- □ Human intransitive adjective
- □ Predicate nouns with Vsup fazer
- □ Predicate nouns with Vsup ser de
- □ PE > PB
- □ PB > PE
- □ SAL
- □ Informal > Formal
- □ Emotions
- ☑ Paraphrasary PT-EN
- □ Paraphrasary PT-BR

Debug □

Input file or text (click to show/hide)

Choose file:

Insert text in the text box

Consideramos que constitui uma provocação quererem arrastar o presidente para uma decisão precipitada.

Results (click to show/hide)

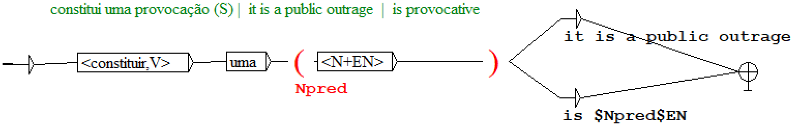
[Consideramos que][**constitui uma provocação**][quererem arrastar o presidente] para uma decisão precipitada .

- is provoking
- it is a public outrage
- is provocative
- Suggest your own paraphrase ›

eSPERTo Paraphrase Generation System

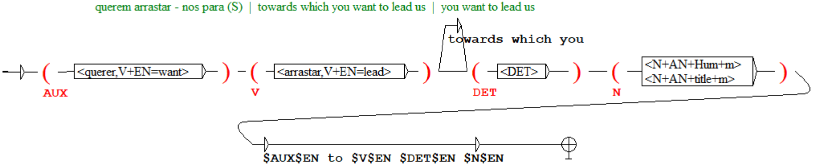
- In the interface image, we illustrate a simple example of using the multilingual paraphrasary to translate multiwords of a sentence in Portuguese into different paraphrases in English
- eSPERTo uses grammars that identify multiwords in a source language, such as *constitui uma provocação* in Portuguese (literally, it 'constitutes a provocation')
- When clicking on this multiword, the text changes to green and the translations of the multiword appear in a drop-down list
- For the Portuguese multiword, eSPERTo shows 3 paraphrases in English
  - is provoking
  - it is a public outrage
  - is provocative
- The suggested translations were paraphrasing pairs in Gold-CLUE and entries in the (EN-PT) Paraphrasary where the same multiword in Portuguese were translations of different multiwords in English

eSPERTo Paraphrase Generation System


constitui uma provocação (S) | it is a public outrage | is provocative
<constituir,V>
uma
(
<N+EN>
)
Npred
it is a public outrage
is $Npred$EN

- The multiword in Portuguese is represented as input of the graph by its constituents: <constituir,V> will match any form of the verb *constituir* in the text, and <N+EN> will match on the predicate noun *provocação*, which will be stored in the variable $Npred
- The top path of the graph will output it is a public outrage whereas the bottom path will output the translation of *provocação* stored in the variable $Npred - *Npred*EN - preceded by *is*

eSPERTo Paraphrase Generation System


querem arrastar - nos para (S) | towards which you want to lead us | you want to lead us
towards which you
<querer,V+EN=want>
AUX
<arrastar,V+EN=lead>
V
<DET>
DET
<N+AN+Hum+m>
<N+AN+title+m>
N
AUXEN to VEN DETEN NEN

- Simplified paraphrasary grammar that allowed for the generation of the distinct translations into English of the multiword [QUERER] *arrastar* [NP+AN+HUM]
- Each multiword constituent will be stored in different variables ($AUX, $V, $DET, and $N) in order to use them to translate them (respectively, $AUX*EN*, $*V*EN, $DET*EN*, *and*$*N*EN))
- The grammar uses the SAL codes +AN, +hum, and +title to restrict the noun in the noun phrase to be human-type
- These grammars take advantage of the multilingual nature of NooJ and other properties included in the dictionary entries
- Full integration of the paraphrasary into the eSPERTo system is under progress

Summary

- Reassessed the concept of alignment, revisited the research on alignments, justified the need for linguistic precision in the alignment task via the analysis and discussion of multiword complexity, crucial in obtaining high quality MT
- Described the Logos Model approach to the processing of multiwords and showed how the SemTab function helped configure and complement our alignment model/proxy
- Presented the Cross-Lingual Unit Elicitation (CLUE) approach based on the CLUE-Guidelines, summarized the CLUE-Aligner tool and the gold collection Gold-CLUE
- The guidelines cover important linguistic phenomena that were left undiscussed in previously presented guidelines
- With a special focus on multiwords, we added an extra level to the alignment process, with the hypothesis that this contributes to a deeper scientific/linguistic process of alignments' annotation

Summary

- The CLUE-Guidelines led to the gold data set Gold-CLUE, which include efficiently-aligned non-contiguous multiwords
- The linguistic analysis undertaken to establish the Gold-CLUE has allowed some advance in the establishment of a standard for the recognition, processing, translation, and evaluation of multiwords
- Some limitation of previous alignment tools (and tasks) motivated the development of the CLUE-Aligner
- All alignments were made in this alignment tool, but only the EN-PT data set was reviewed. We are still in the process of reviewing the other language pairs
- From the EN-PT Gold-CLUE, we selected which entries would go into the multilingual Paraphrasary, either as simple entries or comment lines for rules.
- Illustrate how the collected paraphrases are used in the eSPERTo paraphrase generation system

Future Work

- It is important to develop a more robust resource, with a joint discussion of the most challenging linguistic phenomena of the CLUE-Guidelines to improve areas that are known to be nonconsensual, a more refined methodology, which supports linguistic phenomena in the 4 classes identified in this work
- All data should be multi-annotated by more than two annotators so that no multiword is left unidentified and the coverage of multiword alignments in the data is complete. If done properly inter-annotator agreement should be minimal
- Due to the extent of the work at hand, most linguistic phenomena were left undiscussed
- A detailed analysis of these phenomena is important for the improvement of the alignment techniques and for the enhancement of the quality of MT
- An MT program that offers correct translation of multiwords via paraphrases demonstrates how applied linguistic knowledge helps improve output quality.
- Our goal is the development of an MT model that integrates linguistic knowledge where all sorts of multiwords are included at the alignment level and feed the paraphrasaries that set in motion and enrich the translation engine

Thank you!

inesc id lisboa

TÉCNICO LISBOA