Using a discourse bank and a lexicon for the automatic identification of discourse connectives

Amália Mendes & Iria del Río

CLUL - Centre of Linguistics, University of Lisbon

September 25th, 2018

Outline

Outline

Discourse Analysis and NLP

- POS, lemmatization, or syntactic relations have been consistently addressed for English and other languages with good results in terms of resource availability and tool development
- Work on the **higher levels of text and discourse** is still scarce (even for English)
- Portuguese language: resources and tools for discourse or semantics are few, and are frequently available only for one variety

Resources and NLP Tools for Discourse Analysis

- Corpora with discourse annotations
  - The PDTB style of annotation has been applied to other languages besides English, such as Turkish (METU corpus), Chinese (MCDTB corpus), Czech (PDB), and applied to English and French speech data
  - For Brazilian Portuguese, several corpora have been annotated in the RST and CST frameworks (CSTNews, CorpusTCC, Rhetalho, Summ-it)

Resources and NLP Tools for Discourse Analysis

- Lexicons
  - German: lexicon DiMLex (275 connectives)
  - French: lexicon LEXCONN (328 connectives)
  - Italian: lexicon LiCO (173 connectives)
  - Spanish: DPDE, an online dictionary of Spanish discourse markers (210 entries)

Resources and NLP Tools for Discourse Analysis

- Discourse Parsers
  - Different approaches to discourse parsing, from rule-based methods to machine learning techniques
  - English: RST (HILDA, SPADE), PDTB (Lin et al. 2012)
  - Consistent work for Brazilian Portuguese: the corpora annotated with discourse information have lead to manual and automatic discourse annotation systems in the RST and CST frameworks (RST Toolkit, DiZer, CSTParser)

Outline

TED-MDB Corpus Annotations

- 6 TED talks – English, German, Polish, Portuguese, Russian and Turkish, annotated with PDTB 3.0 format: discourse relations = connectives, arguments and rhetorical sense

I think it's reckless to ignore these things, <u>because</u> **doing so can jeopardize future long-term return.** [Contingency:Cause:reason]

TED-MDB Corpus Annotations

- Explicit and implicit discourse relations
  - **Explicit**: *Ela disse-me que algumas delas não correspondiam à sua marca, às suas expectativas.* Na verdade **uma das obras de tal modo não correspondia à sua marca, que ela tinha-a posto no lixo no seu estúdio.**. (She told me that a few didn't quite meet her own mark for what she wanted them to be. One of the works, in fact, so didn't meet her mark, she had set it out in the trash in her studio)[Expansion:Instantiation] (TED Talk no. 1978)

TED-MDB Corpus Annotations

- Explicit and implicit discourse relations
  - **Implicit**: *esta companhia tem a visão direcionada para o que eles chamam de "o novo Novo Mundo".* (Implicit = porque) **São quatro mil milhões de pessoas da classe média que precisam de comida, de energia e de água.** (this company has their sights set on what they call "the new New World." That's four billion middle class people demanding food, energy and water.) [Contingency:Cause:Reason] (TED Talk no. 1927)

TED-MDB Corpus Annotations

- Inter and intra sentential discourse relations
  - **Inter**: *Eles acreditam que o ASG tem o potencial de criar impacto em riscos e receitas,* assim, **incorporar o ASG no processo de investimento é fundamental ao seu dever de agir no melhor interesse dos membros do fundo...** (They believe that ESG has the potential to impact risks and returns, so incorporating it into the investment process is core to their duty to act in the best interest of fund members...) (TED Talk no. 1927)

TED-MDB Corpus: PDTB rhetorical senses

- PDTB rhetorical senses: hierarchy
  - Top 4 senses: Expansion, Temporal, Contingency, Contrast
  - Subcategories:
    - Expansion:Instantiation
    - Contingency:Cause:Reason

Lexicon of Discourse Markers (LDM-PT)

- Provides a set of lexical items in Portuguese that have the function of structuring discourse and ensuring textual cohesion and coherence at intra-sentential and inter-sentential levels
- Each discourse marker (DM) is associated to the set of its rhetorical senses, following the PDTB 3.0 sense hierarchy
- 252 pairs of DM/rhetorical sense

Discourse connectives in LDM-PT

- Structure discourse - ensure textual cohesion and coherence
- No inflection
- POS: conjunctions, adverbs and adverbial phrases, prepositions
- Single (porque) or multi word (na verdade) units
- Continuous (a fim de) and discontinuous (por um lado... por outro lado) units

LDM-PT Structure

- Pairs of discourse marker - PDTB rhetorical senses
- POS category, internal structure of discourse connective
- Restrictions on the mood and tense of the clause
- English near-synonym(s)

LDM-PT Excel format

/dmarker/@word	dmarker/@	/orths/orth1/@orth1ty	orths/orth1/part1/@orth	dmarker/orths/orth1/orth1part1	rker/orths	dmarker/orths/orth1/orth1part2	dmarker/syn/type	dmarker/syn/cat	syn/context/mood	er/syn/context	dmarker/syn/modifier1	mar	dmarker/sem/relation1	dmarker/sem/relation2	dmarker/sem/relation	
a fim de	dm0	cont	phrasal	a fim de			primary connective	prep		infinitive			contingency	purpose	arg2-as-goal	
a fim de que	dm1	cont	phrasal	a fim de que			primary connective	csu	subjunctive				contingency	purpose	arg2-as-goal	
à medida que	dm2	cont	phrasal	à medida que			primary connective	csu					temporal	synchronous		
a menos que	dm3	cont	phrasal	a menos que			primary connective	csu					comparison	concession	arg2-as-denier	
a não ser que	dm4	cont	phrasal	a não ser que			primary connective	csu					comparison	concession	arg2-as-denier	
a partir do momento em que	dm5	cont	phrasal	a partir do momento em que			secondary connective	csu					temporal	asynchronous	succession	
a propósito de	dm6	cont	phrasal	a propósito de			primary connective	prep					expansion	conjunction		
a razão é que	dm7	cont	phrasal	a razão era			altlex	noun					contingency	cause	reason	

LDM-PT DIMLex format

```
<dmarkers>
    <dmarker word="a fim de que" id="dm1">
            <orth1 type="cont">
                <part1 type="phrasal">a fim de que</part1>
                <part2 type=""></part2>
            </orth1>
        <syn>
            <type>primary connective</type>
            <cat>csu</cat>
            <context>
                <mood>subjunctive</mood>
                <tense></tense>
            </context>
            <modifier1></modifier1>
            <modifier2></modifier2>
        </syn>
        <sem>
            <relation1>contingency</relation1>
            <relation2>purpose</relation2>
            <relation3>arg2-as-goal</relation3>
        </sem>
        <synonym lexicon="dimlex-en" entry-id="22">so that</synonym>
        <examples>
            <example1 source="CRPC">Por fim , a Comissão sugere um sistema de etiquetagem das viaturas a fim de que o cliente possa fazer uma escolha com
            <example2 source=""></example2>
            <example3 source=""></example3>
        </examples>
        <comment></comment>
    </dmarker>
    <dmarker word="à medida que" id="dm2">
            <orth1 type="cont">
                <part1 type="phrasal">à medida que</part1>
                <part2 type=""></part2>
            </orth1>
        <syn>
            <type>primary connective</type>
            <cat>csu</cat>
            <context>
                <mood></mood>
                <tense></tense>
            </context>
            <modifier1></modifier1>
            <modifier2></modifier2>
        </syn>
        <sem>
            <relation1>temporal</relation1>
            <relation2>synchronous</relation2>
            <relation3></relation3>
        </sem>
        <synonym lexicon="dimlex-en" entry-id="28">as</synonym>
        <examples>
            <example1 source="CRPC">À medida que se aproximam as festas de Nossa Senhora da Conceição , sucedem -se as movimentações à sua volta .</examp
            <example2 source=""></example2>
            <example3 source=""></example3>
        </examples>
```

Outline

Automatic Identification of Connectives

- Argument identification is the first step of discourse parsing and has a central role in building quality discourse representations
- We understand argument identification as the identification of the different elements that compose a discourse relation (explicit or implicit and inter or intra- sentential): potential connectives and arguments.
- In this experiment we focused on the identification of connectives

The ambiguity problem...

- In many cases, words that have a cohesive function in texts may also have non connective functions, that is, they are ambiguous:
  1. *Estas iniciativas criam um ambiente de trabalho mais móvel* e **reduzem a nossa pegada imobiliária**. (TED talk 1927) (These initiatives create a more mobile work environment and reduce our housing footprint.)
  2. As companhias *e* os investidores não são os únicos responsáveis pelo destino do planeta . (Companies and investors are not singularly responsible for the fate of the planet.) TED Talk no. 1927
- LDM-PT does not provide any information about connectives ambiguity

Experiment for automatic identification of discourse connectives

- Goal: evaluate which linguistic information is more relevant for the automatic identification of discourse connectives
- Corpus-driven approach: we used data extracted from the TED-MDB pt corpus.
  1. Identification of the ten most common connectives in the corpus
  2. Definition of three levels of linguistic information
  3. Annotation of the corpus with POS and syntactic information
  4. Evaluation of the impact of the levels of information

Identification of target connectives

- As a first step, we extracted all the explicit discourse relations in the corpus and we identified the explicit connectives with their rhetorical sense
  - There are 275 instances of explicit connectives
  - These connectives correspond to 42 different word-forms with 886 cases in the corpus
  - **Therefore, only a 31% of the possible candidates are effectively working as connectives in our data**

Identification of target connectives

- The ten most frequent connectives account for **81% of the total cases**
- They are (by lemma) in the corpus are: *e* (and), *mas* (but), *para* (for/to), *se* (if), *quando* (when), *porque* (because), *depois* (after), *por* (for/because), *ou* (or), *então* (then).

Ten most common connectives

Word-forms	Connectives	NonConnectives
569	224 - 39%	345 -61%

Table 1: Distribution of word-forms, connectives and non-connectives in the corpus for the ten most common connectives.

Annotation of TED-MDB pt corpus

- As a second step, we automatically annotated the PT-TED-MDB corpus with lemma, POS and syntactic information.
  - POS and lemma: Freeling
  - Constituency parsing: PALAVRAS parser

Levels of linguistic information

- To investigate the contribution of different linguistic features to the identification task, we first defined three levels of linguistic information:

1. Word-form of the connective
2. POS + lemma
3. Word-form + POS + lemma + syntactic information involving the connective and its context

- We then applied a rule-based method that makes use of these levels of linguistic information, and we measured precision (and, in some cases, recall) in the identification of connectives and non-connectives in the corpus

Results

Class	P-Conn	P-NConn	R-Conn	R-NConn
`wf`	0.39	0	1	0
`wf+pos+lemma`	**0.41**	1	1	0.09
`wf+pos+lemma+syntax`	**0.85**	0.99	0.99	**0.89**

Table 2: Results for each linguistic level: precision and recall are reported.

Discussion

Word-form

- Each word-form that can be a connective is effectively working as a connective
- Considering that any word that can be a connective is working as such, we obtain a precision of 39% in the identification of connectives and a 0% of precision in the identification of non-connectives (because all occurrences are considered connectives).
- The level of ambiguity changes depending on the connective:
  - *quando* (when) works as a connective in 94% of its occurrences, but has low frequency (6% of the cases)
  - *e* (and) works as a connective in a 37% of the cases, and it is the connective with the highest frequency (37% of the cases)

Discussion

Word-form + POS + Lemma

- Precision improves slightly, from 39% to 41% in the identification of connectives, and from 0% to 100% in the identification of non-connectives
- Recall is 100% for connectives and 9% for non-connectives since, as in the previous approach, we consider most of the candidates as connectives.
- These results make sense considering the fact that connectives are words with low POS ambiguity. Indeed, we can see an improvement for word-forms with more than one POS (that are more or less equally frequent). This is the case of the connective *se* (if), which can be a conjunction (if) or a clitic pronoun.

Discussion

Word-form + lemma + POS + syntax

- Using syntactic information, general precision increases to 85% for connectives and to a 99% in the identification of non-connectives
- Recall of 99% for connectives and a 89% for non-connectives.
- We experiment a slight decrease in recall for connectives and a high increase for non-connectives.

Discussion

Word-form + lemma + POS + syntax

- Syntactic information is especially relevant for connectives that can link different types of structures like conjunctions
- Conjunctions account for a 83.5% of the total connectives in the corpus
  - Remember that the most common connective in the corpus is the copulative conjunction *e*
  - Using syntactic information from PALAVRAS' output, we can identify all the cases where *e* is linking clauses/sentences. Following this approach, we got an 89% of precision and a 100% of recall identifying the connective uses of this conjunction

Discussion

- Connectives that are used in specific constructions could be identified with simpler approaches, like pattern matching - prepositions *por* (because/for) and *para* (for/to)
- Those connectives have a unique POS, and they work as connectives in a very specific construction: when they introduce infinitive subordinated clauses (*para fazer isso* (to do so))
- This simple approach, however, would not be enough for conjunctions like *e* (and) or *mas* (but), that can introduce multiple types of structures and which can be located far from the verb when they introduce clauses. Defining a clause with a surface pattern can be difficult and introduce a lot of errors.

Outline

Now we would like to...

- Extend this approach to all the connectives in our corpus
- Experiment also with a dependency representation
- Explore the identification of arguments and sense attribution for each discourse relation

Thanks!