propor

International Conference on Computational Processing of the Portuguese Language

FINDING OPINION TARGETS IN NEWS COMMENTS AND BOOK REVIEWS

Leonardo G. Catharin and Valéria D. Feltrim

Universidade Estadual de Maringá

Departamento de Informática

September 25, 2018

UEM

Universidade Estadual de Maringá

Introduction

- **Opinion target identification** is an important part of the opinion mining process
  - *What is this opinion about?*

- It usually performed for products' reviews
  → **less common in other domains**

Opinion targets

- **Opinion targets** may be explicit or implicit
- Examples:

"Sócrates é retrógrado e antiquado."

Explicit

"Sensacional!"

Implicit

Purpose

- **We focused on identifying explicit opinion targets** in texts written in Portuguese
- 2 less explored **domains**
  - Comments about political news (SentiCorpus-PT)
  - Reviews about books (Corpus ReLi)
- 3 approaches
  - Centering
  - Pattern matching
  - Heuristics

Corpora

- **SentiCorpus-PT** (Carvalho et al., 2011)
  - **Comments about news articles** covering TV debates on the 2009 election of the Portuguese Parliament
  - ~1,000 comments segmented into ~4,000 sentences
  - Targets are **human entities**
    - Politicians, political organizations (generally used for referring its members), media personalities, or users.
  - 94% of the sentences have annotated targets
    - 79% of them have exactly one target

Corpora

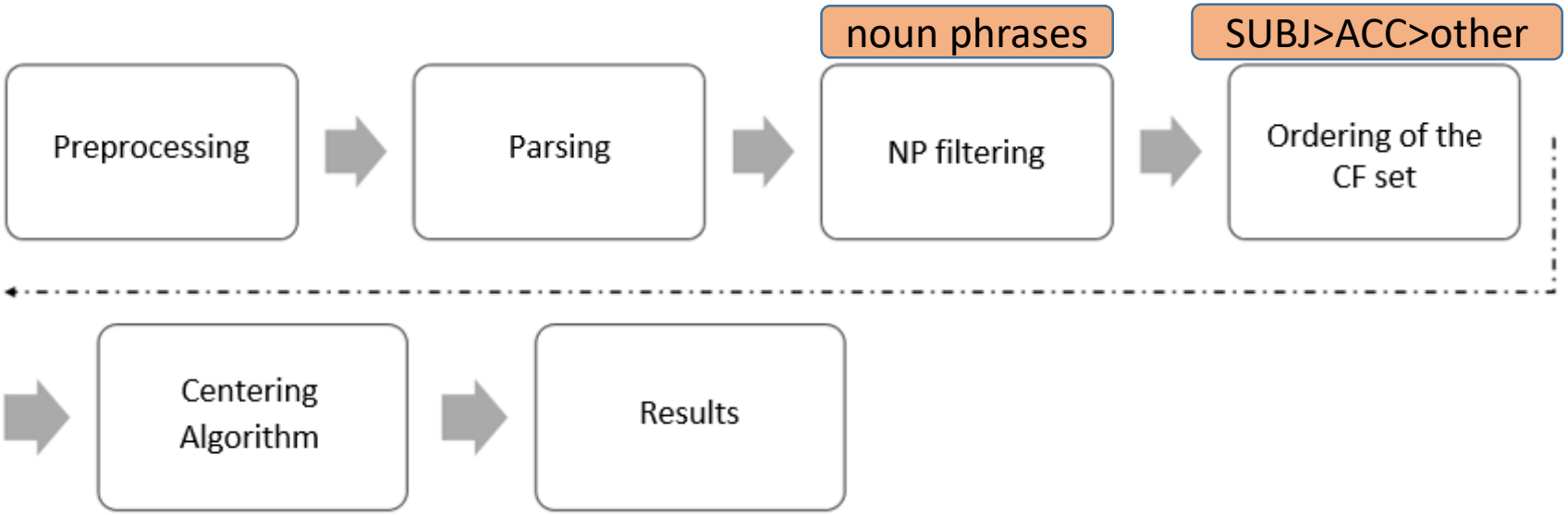
- **ReLi** (FREITAS et al., 2012)
  - **Book reviews** from an on-line social network of readers
  - 1,600 reviews from 13 different books, and it has ~13,000 sentences
  - Targets are either a **book or one of its parts**, such as chapters and characters
  - 18% of the sentences have annotated targets
    - 81% of them have exactly one target
    - A small proportion compared to SentiCorpus-PT

Centering

- **Centering** theory
  - Evaluates coherence in discourse by analyzing transitions among **centers of attention**
    - *"Since the center is the focus of attention, it is likely to be a target"*
- For extracting opinion targets
  - First proposed by Ma and Wan (2010) to identify targets in comments about news articles in Chinese
  - Explicit and implicit targets
    - Besides using CT on the comments, they also used information extracted from the articles

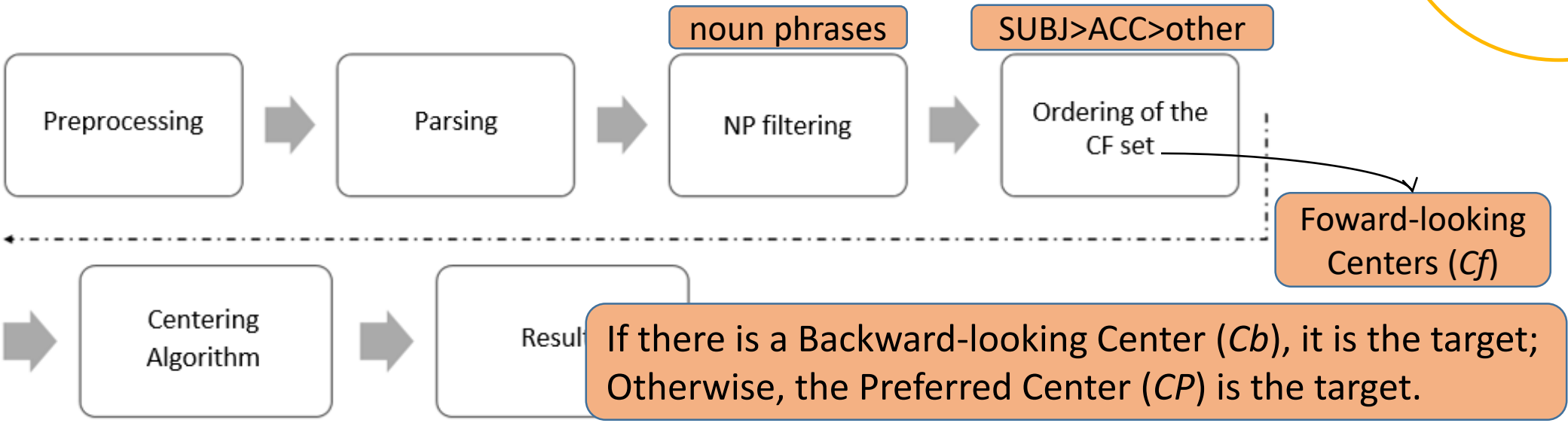
Prototype based on Centering

- Adaptation of Ma and Wan's approach that considers **only explicit targets**
- Since it searches for the most salient center, it **identifies only one target per sentence**



Prototype based on Centering

- Adaptation of Ma and Wan's approach that considers **only explicit targets**
- Since it searches for the most salient center, it **identifies only one target per sentence**

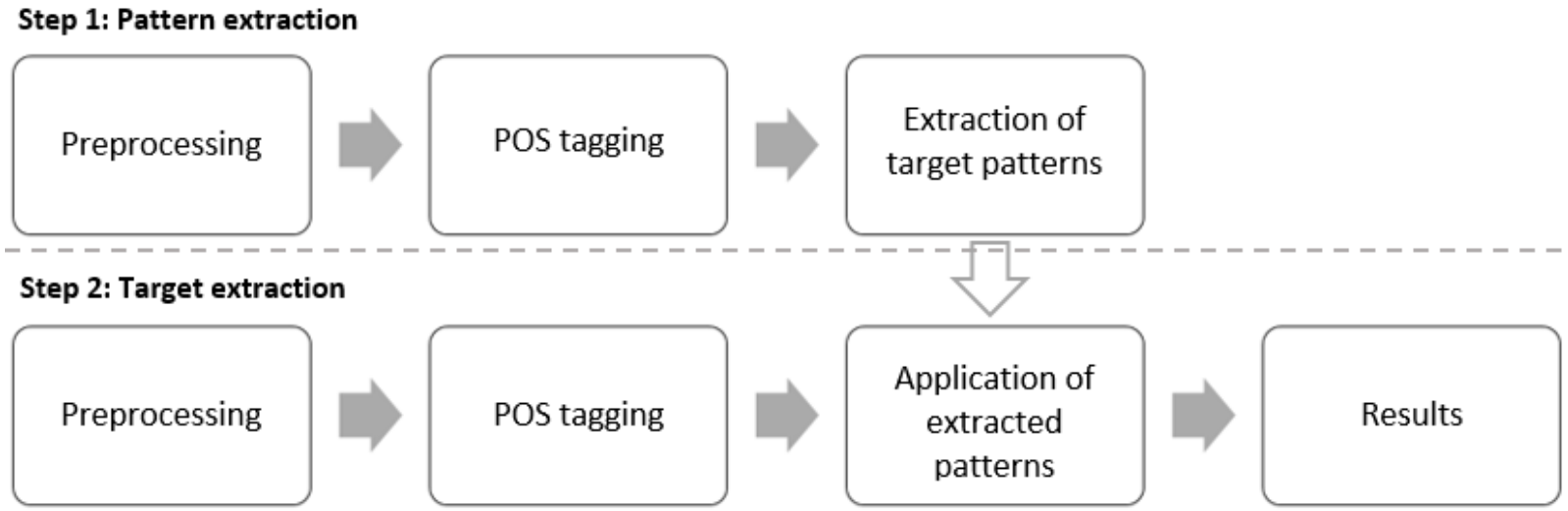


Pattern matching

- **Patterns** are units of information that are recurrent in the text
- For extracting opinion targets
  - Patterns are usually based on POS information
    - Turney (2002), Htay and Lynn (2013), Maharani et al. (2015), Rocha et al. (2016)(named entities)

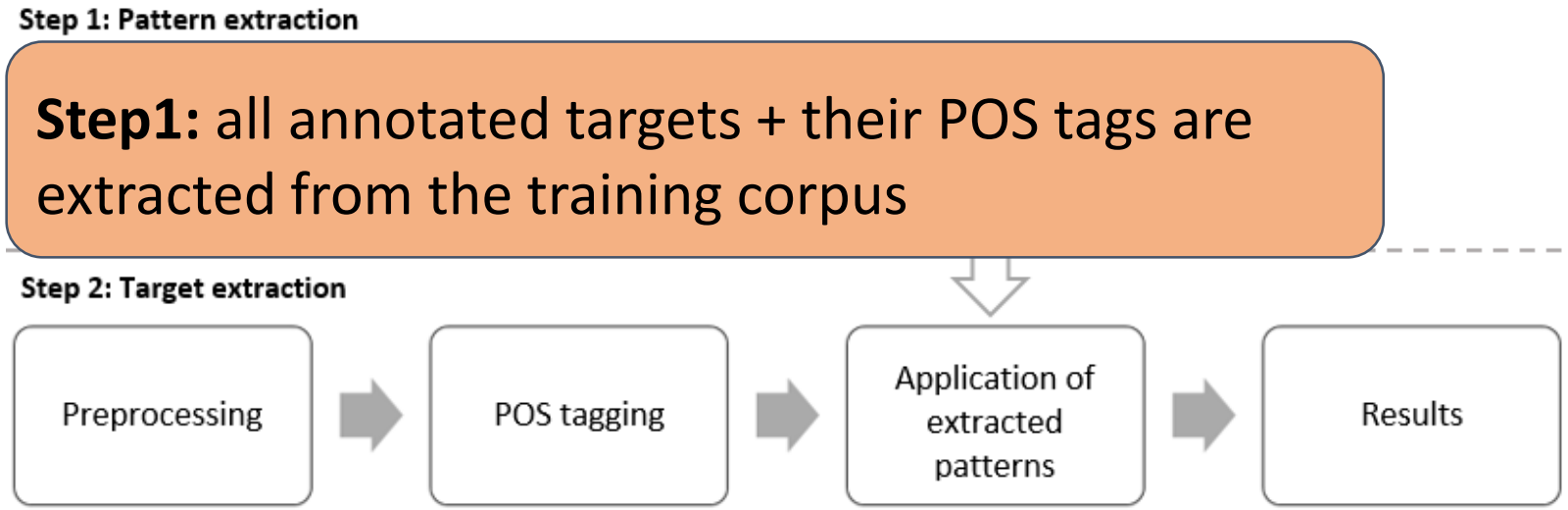
Prototype based on patterns

- A two-step process:
  - Step 1 – pattern extraction: ~ ½ corpus randomly selected for training
  - Step 2 – target extraction: the remaining of the corpus used for testing



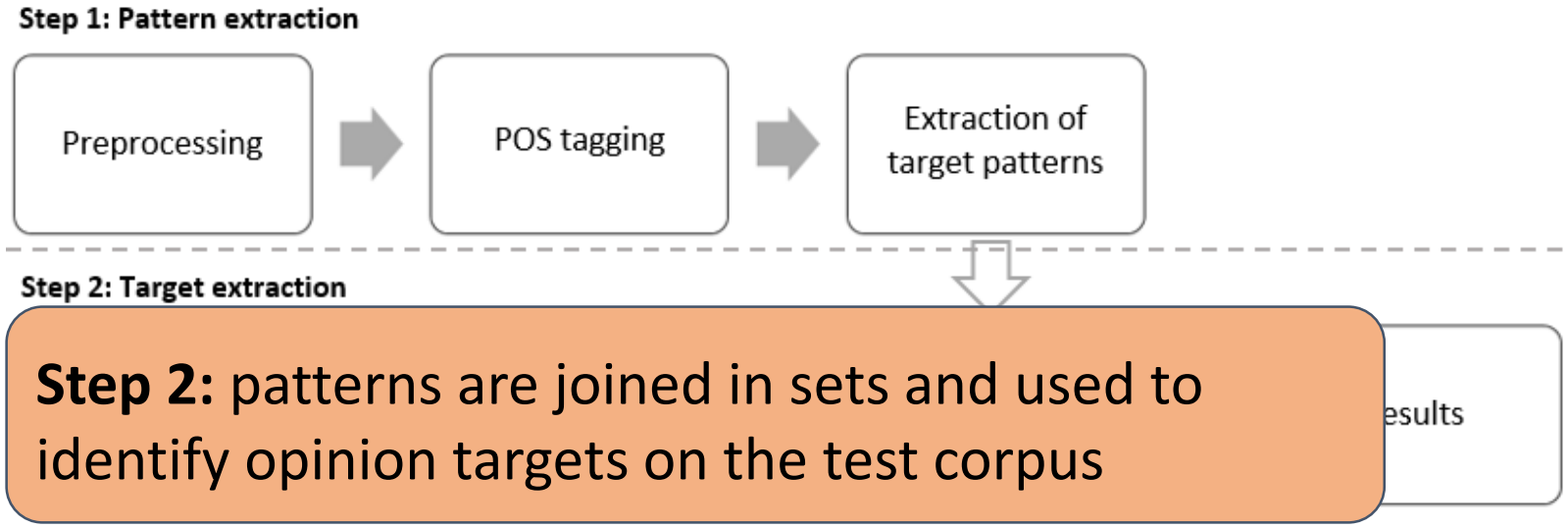
Prototype based on patterns

- A two-step process:
  - Step 1 – pattern extraction: ~ ½ corpus randomly selected for training
  - Step 2 – target extraction: the remaining of the corpus used for testing



Prototype based on patterns

- A two-step process:
  - Step 1 – pattern extraction: ~ ½ corpus randomly selected for training
  - Step 2 – target extraction: the remaining of the corpus used for testing



Step 2: patterns are joined in sets and used to identify opinion targets on the test corpus

Heuristics

- 3 baselines based on the following intuition:
  - *“Many targets are proper names or have the function of subject”*

Baseline 1: the subject of the sentence is the target

If the parser identifies more than one subject, they are ranked by their POS, as follows: proper name > noun > others.

The best ranked subject is the target.

One target per sentence

Heuristics

- 3 baselines based on the following intuition:
  - *“Many targets are proper names or have the function of subject”*

Baseline 1: the subject of the sentence is the target

Baseline 2: the first proper name is the target

Regardless of its syntactic function in the sentence.

One target per sentence

Heuristics

- 3 baselines based on the following intuition:
  - *"Many targets are proper names or have the function of subject"*

Baseline 1: the subject of the sentence is the target

Baseline 2: the first proper name is the target

Baseline 3: all proper names are targets

Regardless of their syntactic function in the sentence.

Many targets per sentence

Results (Centering)

- **SentiCorpus-PT**

Prototype	Precision	Recall	F-Measure
Baseline 1	0.61	0.31	0.41
Baseline 2	0.79	0.38	0.51
CT without correference	0.47	0.29	0.36
CT with correference	0.58	0.36	0.44

- **ReLi**

Prototype	Precision	Recall	F-Measure
Baseline 1	0.11	0.22	0.15
Baseline 2	0.06	0.00	0.06
CT without correference	0.10	0.27	0.15

Results (Pattern matching)

- **SentiCorpus-PT**

Prototype	Precision	Recall	F-Measure
Baseline 3	0.70	0.61	0.65
Set 1 (5-most frequent)	0.25	0.91	0.39
Set 2 (most precise + prop)	0.56	0.63	0.59
Set 3 (all prop)	0.63	0.62	0.63
Set 4 (5 prop + pron-pers)	0.52	0.71	0.60

- **ReLi**
  - ReLi has few targets in relation to its number of sentences, so we could not found patterns that occur with relevant frequency
  - For this reason, this prototype was tested only on SentiCorpus-PT

Conclusion

- We tested 3 approaches for extracting explicit opinion targets in two corpora in Portuguese
- Results were better for SentiCorpus-PT. Why?
  - ReLi has a low ratio of explicit target per sentence
    - The CT prototype usually extracts one target per sentence, so the performance in ReLi was much lower than in SentiCorpus-PT
    - It also impaired the search for frequent patterns, and therefore this approach was not tested on this corpus
  - Targets are mentioned in different ways in the two corpora
    - Targets in SentiCorpus-PT are mostly mentioned by proper names (~60%), which improved the results of the heuristics and the pattern matching
  - Further analysis of ReLi would be necessary to guide the proposal of new approaches and better heuristics for this corpus

References

Carvalho, P., Teixeira, J., Sarmento, L., Silva, M. J.: Liars and Saviors in a Sentiment Annotated Corpus of Comments to Political Debates. In: 49th Annual Meeting of The Association for Computational Linguistics, pp. 564-568. ACL, Portland (2011)

Freitas, C., Motta, E., Milidiu, R. L., Cesar, J.: Sparkling Vampire... lol! Annotating Opinions in a Book Review Corpus. In: Aluísio, S.M., Tagnin, S.E.O. (eds.), New Language Technologies and Linguistic Research: A Two-Way Road, pp. 128-146. Cambridge Scholars Publishing (2014)

Htay, S.S., Lynn, K.T.: Extracting Product Features and Opinion Words Using Pattern Knowledge in Customer Reviews. The Scientific World Journal, 2013, 5p. (2013)

Liu, K., Xu, L., Zhao, J.: Syntactic Patterns versus Word Alignment: Extracting Opinion Targets from Online Reviews. In: 51th Annual Meeting of The Association for Computational Linguistics, pp. 1754-1763. ACL, Sofia (2011)

Ma, T., Wan, X: Opinion target extraction in Chinese news comments. In: 23th International Conference on Computational Linguistics, pp. 782-790. ACL, Beijing (2010)

Maharani, W., Widyantoro, D. H., Khodra, M. L.: Aspect Extraction in Customer Reviews Using Syntactic Pattern. Procedia Computer Science 2(59), 244-253 (2015)

References

Qiu, G. Liu, B., Bu, J., Chen, C.: Opinion word expansion and target extraction through double propagation. Computational Linguistics, 37(1), 9-27. ACL (2011)

Rocha, C., Jorge A. M.; Sionara, R. A, Brito, P., Pimenta, C., Rezende, S. O.: PAMPO: using pattern matching and pos-tagging for effective named entities recognition in Portuguese. ARXIV PREPRINT ARXIV:1612.09535 2(23), (2016)

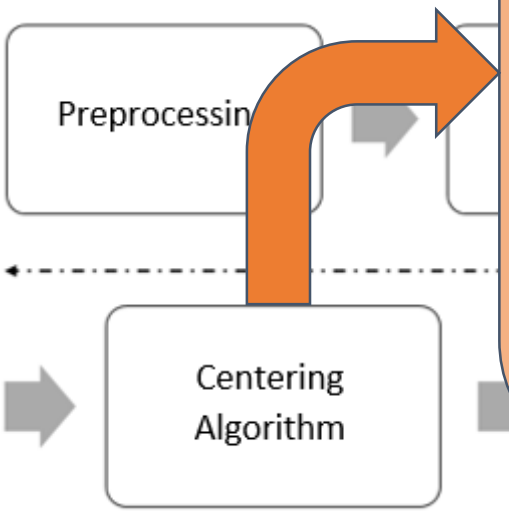
Turney, P. D.:Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: 40th Annual Meeting of The Association for Computational Linguistics, pp. 417-424. ACL, Philadelphia (2002)

Prototype based on Centering

- Adaptation of Ma and Wan's approach that considers **only explicit targets**
- Since it se
only one

For each sentence s_i in text S
  If $i=1$ (s_i is the first sentence)
    Choose the highest ranked element in $Cf(s_i)$ as target t_i
  Else
    For each c_i in $Cf(s_i)$
      If c_i realizes (equals or refers to) an element $c'i$ in $Cf(s_{i-1})$
        Add c'_i to $Cb(s_i)$
    If $Cb(s_i)$ is not empty
      Choose the highest ranked element in $Cb(s_i)$ as t_i
    Else
      Choose the highest ranked element in $Cf(s_i)$ as t_i

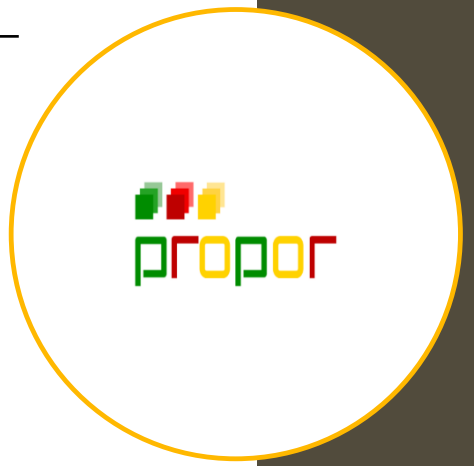
Adapted from Ma and Wan (2010)



Results for pattern matching

- The four best-performing sets of patterns

Set 1 (5-most frequent)	{prop, n, pron-pers, prop prop, prop prop prop}
Set 2 (most precise + prop)	{pron-det prp+pron-pers, prop prp+art prop, n prop prop, n prp n prp+art prop, prp+pron-pers, n prp+art prop, n prp prop prop, prop prp prop, pron-pers pron-det, prop prp+art n, prop prop, prop}
Set 3 (all prop)	{prop, prop prop, prop prop prop, prop prp prop, prop prop prop prop, n prp+art prop, n prop, prop adj, n prop prop, prop prp+art prop, prop prp+art n, n prp n prp+art prop, n prp+art prop prop, n prop prop prop, n prp prop prop, n prp prop}
Set 4 (5 prop + pron-pers)	{pron-det prp+pron-pers, prop prp+art prop, n prop prop, n prp n prp+art prop, prp+pron-pers, n prp+art prop, n prp prop prop, prop prp prop, pron-pers pron-det, prop prp+art n, prop prop, prop, prop prop prop, prop prop prop prop, pron-pers pron-det, pron-pers}



Corpora

- Corpora data summary

SentiCorpus-PT				ReLi			
Sentences with	**#**	**Targets with**	**#**	**Sentences with**	**#**	**Targets with**	**#**
No targets	221	1 word	3331	No targets	10281	1 word	2337
1 target	3071	2 words	597	1 target	1818	2 words	102
2 targets	494	3 words	370	2 targets	320	3 words	108
3 targets	77	4 words	69	3 targets	72	4 words	84
4 targets	18	5 words	20	4 targets	20	5 words	47
5-6 targets	7	6-8 words	13	5 targets	3	6-20 words	89