OVERVIEW OF THE ASSIN SHARED TASK AND CORPUS

Erick Fonseca, Sandra Maria Aluísio, Leandro B. dos Santos , Marcelo Criscuolo

Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo

Tomar, Portugal — PROPOR 2016

INTRODUCTION

OVERVIEW

- **ASSIN**: Avaliação de Similaridade Semântica e Inferência Textual
  - Workshop co-located with PROPOR 2016
  - Shared task on Semantic Similarity and Recognizing Textual Entailment in Portuguese
  - Annotated corpus used in the shared task

OVERVIEW

- **ASSIN**: Avaliação de Similaridade Semântica e Inferência Textual
  - Workshop co-located with PROPOR 2016
  - Shared task on Semantic Similarity and Recognizing Textual Entailment in Portuguese
  - Annotated corpus used in the shared task
- Six participants (good number for Portuguese NLP)
  - Three from Brazil, three from Portugal
  - All took part in STS, but only four in RTE

TASKS DEFINITION

Both tasks deal with sentence pairs

TASKS DEFINITION

Both tasks deal with sentence pairs

Entailment Entailment, paraphrase and neutral

TASKS DEFINITION

Both tasks deal with sentence pairs

Entailment Entailment, paraphrase and neutral
No contradiction class: almost none in the corpus

TASKS DEFINITION

Both tasks deal with sentence pairs

Entailment Entailment, paraphrase and neutral
No contradiction class: almost none in the corpus

Similarity Continuous numeric decision from 1 to 5

SIMILAR INITIATIVES

- RTE Challenges[1]
  - Text entailment; short sentences as the second component
- SICK[2]
  - Entailment and similarity; simple sentences semi-automatically created
- STS at Semeval[3]
  - Similarity; many sources

---

[1] Luisa Bentivogli et al. (2009). "The fifth pascal recognizing textual entailment challenge". In: *Proceedings of the Text Analysis Conference 2009*.

[2] Marco Marelli et al. (2014). "SemEval-2014 Task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment". In: *Proceedings of the 8th International Workshop on Semantic Evaluation*.

[3] Eneko Agirre et al. (2015). "SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability". In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

THE ASSIN CORPUS

DESIDERATA

- We needed a corpus with **related** sentence pairs
  - Varying degrees of similarity
  - Entailment/paraphrase/neutral

DESIDERATA

- We needed a corpus with **related** sentence pairs
  - Varying degrees of similarity
  - Entailment/paraphrase/neutral
- We wanted to avoid trivial decisions
  - Neutral pairs should have similar words and concepts
  - Entailment/paraphrase pairs should not be almost the same

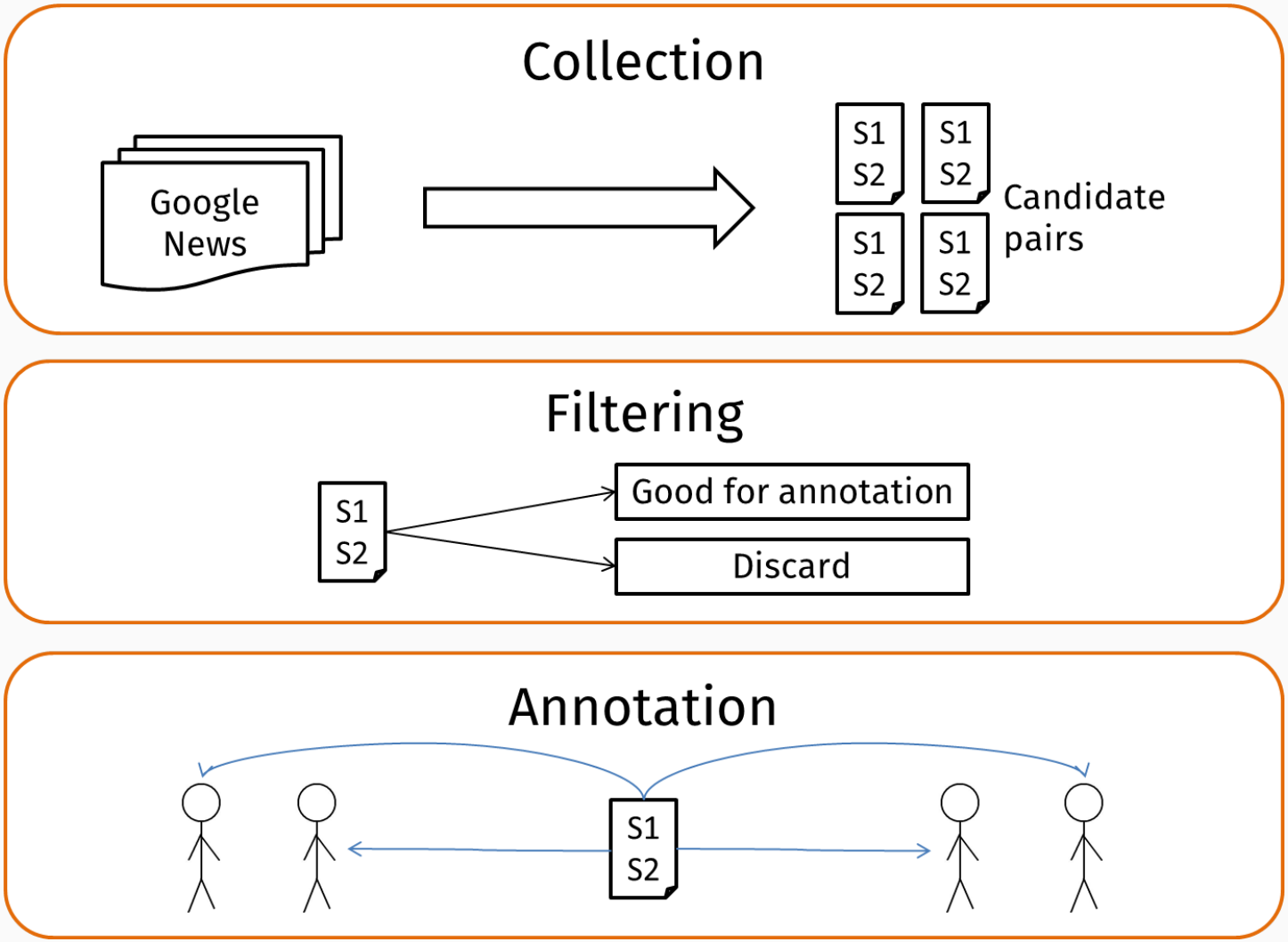

Figure 1: Corpus creation workflow

CANDIDATE COLLECTION

- We used Google News Brazil and Portugal
  - News on the same subject are clustered together

CANDIDATE COLLECTION

- We used Google News Brazil and Portugal
  - News on the same subject are clustered together

- LDA models select similar sentences within each cluster

CANDIDATE COLLECTION

- We used Google News Brazil and Portugal
  - News on the same subject are clustered together
- LDA models select similar sentences within each cluster
- Tunable parameters
  - Similarity score (0.6 – 0.9)
  - Minimum and maximum exclusive words (0.1 – 0.8)
  - Sentence size (up to 25 tokens, except stopwords)

PAIR FILTERING

- Pairs were manually checked

PAIR FILTERING

- Pairs were manually checked
- Pairs with irrelevant sentences were discarded
  - Game scores, currency values, wrong sentence splits, meaningless sentence without context etc.

PAIR FILTERING

- Pairs were manually checked
- Pairs with irrelevant sentences were discarded
  - Game scores, currency values, wrong sentence splits, meaningless sentence without context etc.
- Minor corrections on spelling and encoding

PAIR FILTERING

- Pairs were manually checked
- Pairs with irrelevant sentences were discarded
  - Game scores, currency values, wrong sentence splits, meaningless sentence without context etc.
- Minor corrections on spelling and encoding
- Editions to favor entailment and paraphrase
  - Previous experiments showed these are rare
  - Adding/removing information such as date, time, purpose, secondary clauses

ANNOTATION

- Each pair was annotated by four people
  - Web interface
  - Annotators assigned randomly to each pair

ANNOTATION

- Each pair was annotated by four people
  - Web interface
  - Annotators assigned randomly to each pair
- 36 people annotated different quantities of the dataset
  - All annotators were trained and had 18 examples to practice with

ANNOTATION

- Each pair was annotated by four people
  - Web interface
  - Annotators assigned randomly to each pair
- 36 people annotated different quantities of the dataset
  - All annotators were trained and had 18 examples to practice with
- Whole annotation took around 6 months

GUIDELINES

- Some guidelines were developed to standardize annotation
    - Reduce subjectivity

GUIDELINES

- Some guidelines were developed to standardize annotation
    - Reduce subjectivity
- Not many rules:

GUIDELINES

- Some guidelines were developed to standardize annotation
  - Reduce subjectivity
- Not many rules:
  - Ignore the current date (*in 1945* vs. *70 years ago*)

GUIDELINES

- Some guidelines were developed to standardize annotation
  - Reduce subjectivity
- Not many rules:
  - Ignore the current date (*in 1945* vs. *70 years ago*)
  - Named entities are the same even if one has an appositive (*Brazilian newspaper Folha* vs *Folha*)

GUIDELINES

- Some guidelines were developed to standardize annotation
  - Reduce subjectivity
- Not many rules:
  - Ignore the current date (*in 1945* vs. *70 years ago*)
  - Named entities are the same even if one has an appositive (*Brazilian newspaper Folha* vs *Folha*)
  - Indirect speech entails the statement, but not the opposite

GUIDELINES

- Some guidelines were developed to standardize annotation
  - Reduce subjectivity
- Not many rules:
  - Ignore the current date (*in 1945* vs. *70 years ago*)
  - Named entities are the same even if one has an appositive (*Brazilian newspaper Folha* vs *Folha*)
  - Indirect speech entails the statement, but not the opposite
  - Numbers must match exactly

FINAL DATASET

- The final dataset has 10,000 pairs
  - 5,000 Brazilian Portuguese and 5,000 European Portuguese

FINAL DATASET

- The final dataset has 10,000 pairs
  - 5,000 Brazilian Portuguese and 5,000 European Portuguese
- Only pairs with 3 or more agreeing entailment judgments were kept
  - 11.3% of the total of annotated pairs were discarded

FINAL DATASET

- The final dataset has 10,000 pairs
  - 5,000 Brazilian Portuguese and 5,000 European Portuguese
- Only pairs with 3 or more agreeing entailment judgments were kept
  - 11.3% of the total of annotated pairs were discarded
- Similarity averaged over 4 scores
  - Final values have intervals of 0.25

Metric	Value
Pearson correlation	0,74
Mean Standard Deviation	0,49
Fleiss's κ	0,61
Concordance	0,80

Table 1: Annotation Statistics

Similarity	PB	PE	Total
4,0 – 5,00	1.074	1.336	2.410
3,0 – 3,75	1.591	1.281	2.872
2,0 – 2,75	1.986	1.828	3.814
1,0 – 1,75	349	555	904
Mean	3,05	3,05	3,05

Table 2: Similarity statistics

Relation	BP	EP	Total
Neutral	3.884	3.432	7.316
Entailment	870	1.210	2.080
Paraphrase	246	358	604

Table 3: Entailment statistics

EVALUATION

METRICS

Similarity Pearson correlation and mean squared error (MSE)

METRICS

Similarity Pearson correlation and mean squared error (MSE)

Entailment F1 and accuracy

BASELINES

Two simple baselines were tried

BASELINES

Two simple baselines were tried

Majority Always answer the majority entailment class and the similarity average on the training corpus

BASELINES

Two simple baselines were tried

Majority Always answer the majority entailment class and the similarity average on the training corpus

Overlap Train a logistic regression classifier and a linear regressor with 2 features: proportion of words exclusive to each sentence

SIMILARITY RESULTS

		BP		EP		Total	
Team	Run	Pearson	MSE	Pearson	MSE	Pearson	MSE
Solo Queue	1	0,58	0,50	0,55	0,83	0,56	0,66
	2	0,68	0,41	0,00	1,55	0,29	0,98
	3	**0,70**	**0,38**	0,70	0,66	**0,68**	**0,52**
Reciclagem	1	0,59	1,36	0,54	1,10	0,53	1,23
	2	0,59	1,31	0,53	1,14	0,54	1,23
	3	0,58	1,37	0,53	1,18	0,53	1,27
Blue Man Group	1	0,65	0,44	0,63	0,73	0,63	0,59
	2	0,64	0,45	0,64	0,72	0,63	0,59
ASAPP	1	0,65	0,44	0,68	0,70	0,65	0,57
	2	0,65	0,44	0,67	0,71	0,64	0,58
	3	0,65	0,44	0,68	0,73	0,65	0,58
LEC-UNIFOR	1	0,62	0,47	0,64	0,72	0,62	0,59
	2	0,56	2,83	0,59	2,49	0,57	2,66
	3	0,61	1,29	0,63	1,04	0,61	1,17
L2F/INESC-ID	1			**0,73**	**0,61**		
	2			0,63	0,70		
	3			0,63	0,70		
Baseline (average)	–	0,00	0,76	0,00	1,19	-0,08	0,97
Baseline (overlap)	–	0,63	0,46	0,64	0,75	0,62	0,60

ENTAILMENT RESULTS

		BP		EP		Total	
Team	Run	Accuracy	F1	Accuracy	F1	Accuracy	F1
Reciclagem	1	77,65%	0,29	73,10%	0,43	75,38%	0,4
	2	79,05%	0,39	72,10%	0,38	75,58%	0,38
	3	78,30%	0,33	70,80%	0,32	74,55%	0,32
Blue Man Group	2	81,65%	0,52	77,60%	0,61	79,62%	0,58
ASAPP	1	81,20%	0,50	77,75%	0,57	79,47%	0,54
	2	81,65%	0,47	78,90%	0,58	80,27%	0,54
	3	77,10%	0,50	74,35%	0,59	75,72%	0,55
L2F/INESC-ID	1			**83,85%**	**0,7**		
	2			78,50%	0,58		
	3			78,50%	0,58		
Baseline (majority)	–	77,65%	0,29	69.30%	0,27	73,47%	0,28
Baseline (overlap)	–	**82,80%**	**0,64**	81,75%	**0,7**	**82,27%**	**0,67**

COMMENTS

- Both baselines and evaluation script available at `http://github.com/erickrf/assin`
- Average baseline is easily outperformed, but the overlap one is competitive!
  - Especially in RTE

COMMENTS

- Both baselines and evaluation script available at `http://github.com/erickrf/assin`
- Average baseline is easily outperformed, but the overlap one is competitive!
  - Especially in RTE
- This was unexpected – all participants had sound strategies

COMMENTS

- Both baselines and evaluation script available at `http://github.com/erickrf/assin`
- Average baseline is easily outperformed, but the overlap one is competitive!
  - Especially in RTE
- This was unexpected – all participants had sound strategies
- It means that both tasks are strongly related to word overlap
  - Even though we tried to minimize it during corpus creation

CONCLUSIONS

CORPUS CREATION

- Our method worked, but has a few difficulties...

CORPUS CREATION

- Our method worked, but has a few difficulties...
- Pair filtering/editing is a bottleneck
  - Required knowledge of the task
  - Can’t be done via crowdsourcing

CORPUS CREATION

- Our method worked, but has a few difficulties...
- Pair filtering/editing is a bottleneck
  - Required knowledge of the task
  - Can't be done via crowdsourcing
- The annotation (especially RTE) is subjective
  - Even with the filtering stage and guidelines

APPROACHES

- Participants explored very different strategies

APPROACHES

- Participants explored very different strategies

- Very simple approaches achieved some of the best results

APPROACHES

- Participants explored very different strategies
- Very simple approaches achieved some of the best results
- No participants explored syntactic or semantic structure

APPROACHES

- Participants explored very different strategies
- Very simple approaches achieved some of the best results
- No participants explored syntactic or semantic structure
- Only one team tried deep neural networks, but had bad results

FUTURE WORK

- A future edition of ASSIN would be interesting
  - RTE and STS in Portuguese have much to improve!

FUTURE WORK

- A future edition of ASSIN would be interesting
  - RTE and STS in Portuguese have much to improve!
- ... but the nature of the corpus should be different, especially for RTE
  - Either with simpler sentences, like SICK
  - or with simple facts as the hypotheses

FUTURE WORK

- A future edition of ASSIN would be interesting
  - RTE and STS in Portuguese have much to improve!
- ... but the nature of the corpus should be different, especially for RTE
  - Either with simpler sentences, like SICK
  - or with simple facts as the hypotheses
- We need **less** subjectivity!

Questions?