Computational Linguistics: Part III: NLP applications: Entailment

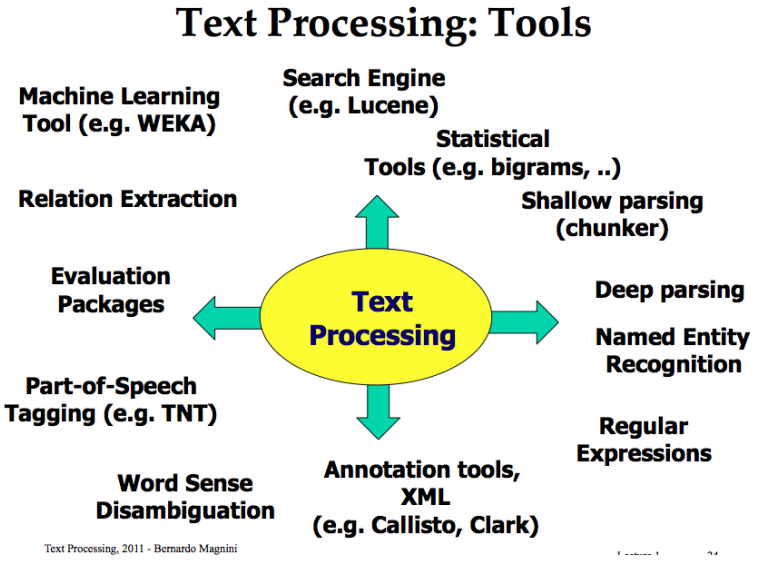
RAFFAELLA BERNARDI

UNIVERSITÀ DEGLI STUDI DI TRENTO

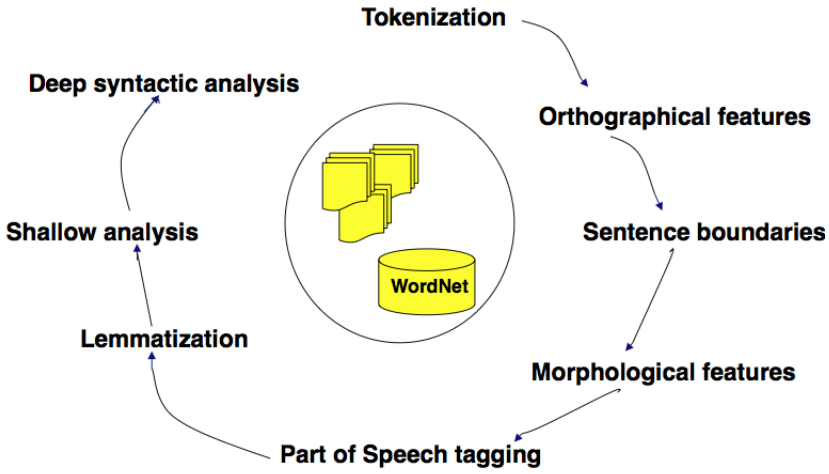
E-MAIL: BERNARDI@DISI.UNITN.IT

Contents

1. NLP tools


Text Processing: Tools
Machine Learning Tool (e.g. WEKA)
Search Engine (e.g. Lucene)
Statistical Tools (e.g. bigrams, ..)
Relation Extraction
Shallow parsing (chunker)
Evaluation Packages
Text Processing
Deep parsing
Named Entity Recognition
Part-of-Speech Tagging (e.g. TNT)
Regular Expressions
Word Sense Disambiguation
Annotation tools, XML (e.g. Callisto, Clark)
Text Processing, 2011 - Bernardo Magnini

1.1. NLP pipe line



2. NLP applications

What we have seen so far has lead to the development of several NLP tools which can be used either alone or (mostly) together as part of complex systems that are able to tackle some tasks. For instance:

- Given a query, they retrieve relevant document IR
- Given a question, they provide the answer QA

Today, we will look at a sub-task behind both IR and QA, viz. Textual Entailment. Tomorrow, we will look at IR and QA.

3. Logical Entailment

A set of premises entails a sentence

$$\{P_1, \ldots, P_n\} \models C$$

if the conclusion is true in every circumstance (possible worlds) in which the premises are true.

When this condition is met, the entailment is said to be *valid*.

Formal Semantics approaches to the entailment would require:

1. natural language sentences to be translated into a Logical Language (mostly FoL)
2. a theorem prover or a model builder to verify whether the entailment is valid.

4. Natural Logic

Natural logic: a logic whose vehicle of inference is natural language. (Suppes 1979, Van Benthem 1986 etc.)

Research question: study how natural language structures contribute to natural reasoning.

$$\frac{\text{Everybody } (\textit{left something expensive})^{+}}{\text{Everybody } (\textit{left something})} \qquad \frac{\text{Nobody } (\textit{left yet})^{-}}{\text{Nobody } \textit{left in a hurry yet}}$$

$$\frac{\text{Not every } (\textit{good logician})^{+} \text{ wonders}}{\text{Not every } \textit{logician} \text{ wonders}} \qquad \frac{\text{Every } (\textit{logician})^{-} \text{ wonders}}{\text{Every } \textit{good logician} \text{ wonders}}$$

4.1. Natural Logic system

MacCartney:

> "FoL and theorem prover or model builder are precise but brittle. Difficult to translate natural language sentences into FoL.

Many inferences are outside the scope of natural logic still a natural logic system can be designed to integrate with other kinds of reasoners.

Natural Logic in NLP: http://nlp.stanford.edu/projects/natlog.shtml

4.2. FraCaS data set

http://www-nlp.stanford.edu/~wcmac/downloads/fracas.xml

Inferences based on Generalized Quantifiers, Plurals, Anaphora, Ellipsis, Comparatives, Temporal References, etc. Eg. GQ's Properties:

Conservativity Q As are Bs == Q As are As who are Bs

- P1 An Italian became the world's greatest tenor.
- Q Was there an Italian who became the world's greatest tenor?

Monotonicity Q As are Bs and all Bs are Cs, then Q As are Cs

- P1 All Europeans have the right to live in Europe.
- P2 Every European is a person.
- P3 Every person who has the right to live in Europe can travel freely within Europe.
- Q Can all Europeans travel freely within Europe?

5. Recognize Textual Entailment: evaluation data sets

Recognizing Textual Entailment (RTE) an International campaign on entailment.

- Started in 2005. (Magnini – FBK – among the first organizers.)
- Data Sets: PASCAL Recognizing Textual Entailment (RTE) challenges.
- Goal: check whether one piece of text can *plausibly* be inferred from another. The truth of the hypothesis is highly plausible, for most practical purposes, rather than certain.

  T ENTAILS H IF, *TYPICALLY*, A HUMAN READING T WOULD INFER THAT H IS *MOST LIKELY* TRUE

T (Text) are *fragments* of text.

RTE-1: http://pascallin.ecs.soton.ac.uk/Challenges/RTE/Introduction/

5.1. RTE 1 examples

T: Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.
H: Yahoo bought Overture
TRUE

T: The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.
H: Israel was established in May 1971
FALSE

T: Since its formation in 1948, Israel fought many wars with neighboring Arab countries.
H: Israel was established in 1948
TRUE

5.2. RTE challenges

- RTE-1 (2005)
- RTE-2
- RTE-3 longer texts (up to one paragraph).
- RTE1-RTE3: entailed vs. not-entailed.
- RTE 4: entailed vs. contradiction (the negation of H is entailed from T) vs. unknown.
- ...

Applied semantic inference. Data sets collected from NLP application scenarios: etc, QA, IR, IE,

Evaluation measures Accuracy percentage of pairs correctly judged and Average precision: ranking based on the system's confidence.

5.3. Data sets: Which (semantic) challenge?

How far can we go just with a parser?

- RTE-1: 37% of the test items can be handled by syntax. 49% of the test item can be handled by syntax plus lexical thesaurus. Syntax good for “true”, less for “false”.
- In RTE-2 65.75% involves deep reasoning.
- RTE-3 data set: Clark et al. imp common understanding of lexical and world knowledge.

The traditional RTE main task, carried out in the first five RTE challenges, consisted of making entailment judgments over isolated T-H pairs. In such a framework, both Text and Hypothesis were artificially created in a way that they did not contain any references to information outside the T-H pair. As a consequence, the context necessary to judge the entailment relation was given by T, and only language and world knowledge were needed, while reference knowledge was typically not required.

5.4. More natural scenarios: Entailment within a corpus

RTE 6 emphasised summarization application. Plus entailment within a corpus – more natural scenario.

Given a corpus, a hypothesis H, and a set of "candidate" sentences retrieved by an IR system from that corpus, RTE systems are required to identify all the sentences that entail H among the candidate sentences.

In such a scenario, both T and H are to be *interpreted in the context of the corpus*, as they rely on explicit and implicit references to entities, events, dates, places, situations, etc. pertaining to the topic.

RTE 7 Plus subtask. To judge whether the information contained in each H is *novel* with respect to (i.e., not entailed by) the information contained in the corpus. If entailing sentences are found for a given H, it means that the content of the H is not new.

6. RTE: Approaches

NLP tools (tokenization, PoS, deep parsing, NER, WSD) and lexical resources (WordNet, DIRT, VerbNet, Reuters corpus, English Gigaword, InfoMap, etc) for lexical similarity judgements.

Approaches

- lexical-syntactic and semantic features
- transformations rules
- deep analysis and semantic inference (logical inference and ontology-based techniques) combined with ML system.
- Voting systems.

The main assumption underlying most of the work in this direction is that decomposing the complex entailment problem would improve the performance of RTE systems.

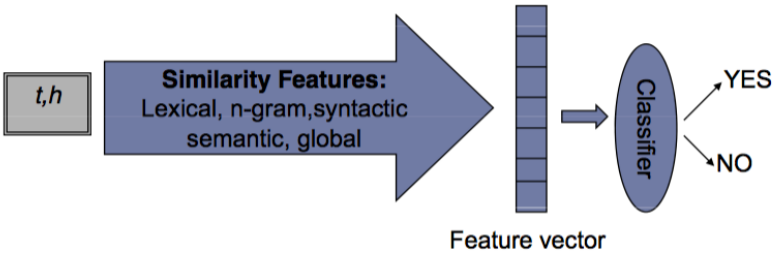
6.1. Classification task

The problem has been seen as a classification task, where features are extracted from the training examples and then used by machine learning algorithms in order to build a classifier, which is finally applied to the test data to classify each pair as either positive or negative

A variety of features has been used, including lexical-syntactic and semantic features, based on document co-occurrence counts, first-order syntactic rewrite rules, and to extract the information gain provided by lexical measures.

E.g Zanzotto, Pennacchiotti and Moschitti 2007

Alignment-based approaches Seeks to refine the similarity approach by defining a meaningful way of determining local similarities between parts of the H and parts of the T, and using the resulting alignment as the basis of a decision function for determining the entailment label



- Features model similarity and mismatch
- Classifier determines relative weights of information sources
- Train on development set and auxiliary *t-h* corpora

6.2. Transformations rules

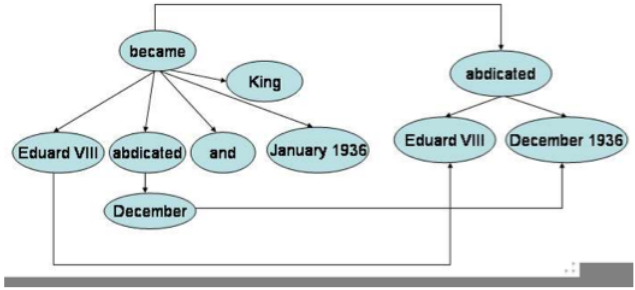
E.g. *Tree Edit Distance* (Kouleykov and Magnini 2005): T entails H if there exists a sequence of transformations applied to T such that we can obtain H with an overall cost below a certain threshold.

The kind of transformations that can be applied (i.e. deletion, insertion and substitution) are determined by a set of predefined entailment rules, which also determine a cost for each edit operation

insertion its cost is proportional to the relevance of the word w to be inserted (i.e. inserting an informative word has an higher cost than inserting a less informative word). (Frequency: The most frequent words (e.g. stop words) have a zero cost of insertion. Position in the syntactic tree.)

substitution its cost is proportional to the entailment relation between the two words: The more the two words are entailed, the less the cost of substituting one word with the other. (Based on WordNet)

deletion Done after alignment.


became
King
abdicated
Eduard VIII
abdicated
and
January 1936
Eduard VIII
December 1936
December

T: Edward VIII became King in January of 1936 and abdicated in December. H: King Ed- ward VIII abdicated in December 1936.

6.3. Deep analysis combined with ML systems

derive a logical representation of T and H. Classical formal model that uses predicate calculus and theorem-proving techniques

(Bos and Markert 2006): Used CCG parser, FoL representations. First order theorem proving and finite model building.

Evaluation: simple word overlap system performs better.

Bos' claim: "There is a place for logic in RTE, but it is (still) overshadowed by the knowledge acquisition problem." Johan Bos, LILT 2014.

6.4. Voting systems

Combine different systems, take their results and choose the result most voted.

7. Alternatives to RTE data sets

- Logical words?
- Pragmatics?
- Composition?

7.1. From RTE to Logic

Dutch Project: “Between Logic and Common Sense: The Formal Semantics of Words” (Winter et al.)

http://logiccommonsense.wp.hum.uu.nl/

1. Phenomena that are commonly involved in entailments.
2. Phenomena that are well understood in the semantic literature and that lend themselves readily to linguistic intuitions as well as to an analysis that is likely to yield high annotation consistency.
3. Phenomena that do not require sophisticated abstract representations and which therefore are easy to classify.

They analyzed RTE 1-3 corpora found out that 77,68% of the entailments are due to one of the following phenomena: restrictive, appositive, and conjunctive modification. Focus on those.

7.2. Restrictive, Appositive and Conjunctive modifications: Examples

- Restrictive
  - T A Cuban American who is accused of espionage pleads innocent.
  - H American accused of espionage
- Appositive
  - T Mr. Conway, Iamgold's chief executive officer, said the vote would be close.
  - H Mr. Conway said the vote would be close.
- Conjunctive
  - T Nixon was impeached and became the first president ever to resign on August 9th 1974
  - Nixon was the first president ever to resign

7.3. RTE extended with the Pragmatics view

Zaenen, Karttunen, Crouch (2005): RTE data sets should be extended so to include:

Entailments due to monotonicity or to temporal and spatial relations.

Conventional implicatures (presuppositions) Facts that are not considered to be part of what makes a sentence true, but the speaker/author is committed to them.

E.g. “Bill acknowledges that the earth is round”

The speaker is committed to the belief that the earth is round.

(a) Kerry realized that Bush was right. & (b) Kerry didn’t realized that Bush was right.

In both (a) and (b) Bush was right.

Conversational Implicatures A collaborative speaker will say as much as she knows. But this implicatures can be cancelled:

1. I had the time to read your paper.
2. CI: I read your paper.
3. I had the time to read your paper, but I decided to go play tennis.

8. Compositional Knowledge

All the data sets above need many NLP tools. How do we evaluate only the Compostional Model?

Sentences Involving Compositional Knowledge (SICK): A data set tailored on CDSMs challenges:

http://alt.qcri.org/semeval2014/task1/

SICK consists of simple sentences that parsers should be able to parse with no mistakes. It does not contain ambiguous sentences, rare words, no named entities, etc.

8.1. How dataset collecation

Starting from the

- 8K ImageFlickr data set: each image is associated with 5 descriptions. We randomly chose 750 images, sampled 2 descriptions from each.
- SemEval-2012 STS MSR-Video descriptions data sets: sentences pairs sampled from the short video snippets. We randomly chose 750 pairs.

These 1500 sentence pairs:

1. normalization: eliminate phenomena outside current CDSM (named entities, nr, multiwords, etc.)
2. expansion: to get sentences with (a) similar; (b) contrasting; (c) different meaning.
3. pairing: each original normalized sentence paired with all other sentences generated from it or from its paired sentence.

8.2. Task: Entailment

Entailment task: Entailment, Contradictin, Neutral?

A Two teams are competing in a football match B Two groups of people are playing football ENTAILMENT
A The brown horse is near a red barrel at the rodeo B The brown horse is far from a red barrel at the rodeo CONTRADICTION
A man in a back jacket is doing tricks on a motorbike A person is riding the bicycle on one wheel NEUTRAL

8.3. Task: Relatedness

Relatedness: 1 to 5?

A A man is jumping into an empty pool
B There is no biker jumping in the air
1.6

A Two children are lying in the snow and are making snow angels
B Two angels are making snow on the lying children
2.9
etc..

8.4. How annotation: Crowdflower

Crowdsources: the process of getting work, usually online, from a crowd of people. Combination of “crowd” and “outsourcing”.

Crowdflower: http://www.crowdflower.com/ To collect and label data.

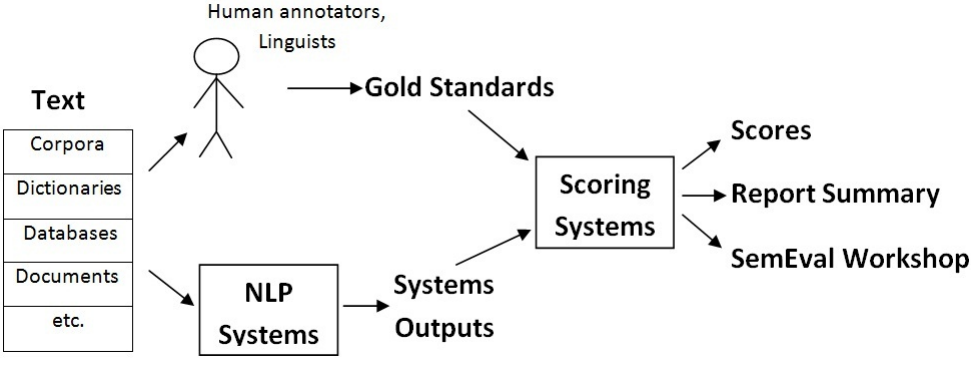
We asked annotators to label the dataset.

- 10 annotators per pair.
- Entailment: majority vote schema
- Relatedness: average of the 10 ratings.

8.5. SemEval: evaluation champaign

Semantic Evaluation Exercises. International Workshop on Semantic Evaluation: http://en.wikipedia.org/wiki/SemEval

First in 1998 (SenSeval till 2004).



8.6. Training, Development, Testing datasets

Participants are provided with:

- Training dataset: To train their systems
- Development data set: To evaluate their system and then improve it.
- Testing dataset: official evaluation phase. Results are submitted. Organizers compute results (compare against gold-standard) and public the results.

8.7. Participants

	Non Composition Features									Comp Features		Learning Methods							External Resources						
Participant ID	Vector Semantics Model	Topic Model	Neural Language Model	Denotational Model	Word Overlap	Word Similarity	Syntactic Features	Sentence difference	Negation Features	Sentence Composition	Phrase composition	SVM and Kernel methods	K-Nearest Neighbours	Classifier Combination	Random Forest	FoL/Probabilistic FoL	Curriculum based learning	Other	WordNet	Paraphrases DB	Other Corpora	ImageFlicker	STS MSR-Video Description	Relatedness Task Ranking	Entailment Task Ranking
ASAP	R	R				R	R	R	R		R			R					R					15	-
ASJAI	B		B	B	B	B	B	B		B		E		B				R		B				17	15
BUAP	B		B			B	B		E		B	E							B					12	7
UEdinburgh	B				B		B		B	B		E	R								B			-	9
CECL	B				B	B		B										B			B			6	6
ECNU	B		B		B	B	B	B		B		B	B	B	B				B		B			1	2
FBK-TR		R			R	R	E	B	E	E	B	R		E					R		R		E	11	11
haLF	E					E				E		E												-	16
IITK	B				B	B	B	B	B		B	B							B					-	-
Illinois-LH	B			B	B	B	B		B	B	B							B	B			B	B	5	1
RTM-DCU	B						B					B			B						B			8	17
SemantiKLUE	B				B	B	B		B			B							B		B			7	4
StandfordNLP	B		B		R	R			R	B							E							2	12
The Meaning Factory	R		R	R	R		R	R		B		E			R	E			B	B	R			3	5
UANLPCourse	B				B	B				B		B												13	18
UIO-Lien									E										E					-	10
UNAL-NLP						B	B		B									B	R	B	B			4	3
UoW				B	B		B		B			B								B				10	8
UQeRsearch					R	R	R	R	R									R	R					14	-
UTexas	B			B					B	B	B					B					B			9	13
Yamarj	B				B	B						B												16	14

8.8. Participating systems: quantitative analysis (Entailment)

ID	Composition	Accuracy (%)
Illinois-LH_run1	P/S	84.6%
ECNU_run1	S	83.6%
UNAL-NLP_run1		83.1%
SemantiKLUE_run1		82.3%
The_Meaning_Factory_run1	S	81.6%*
CECL_ALL_run1		80.0%
BUAP_run1	P	79.7%
UoW_run1		78.5%
Uedinburgh_run1	S	77.1%
UIO-Lien_run1		77.0%
FBK-TR_run3	P	75.4%
StanfordNLP_run5	S	74.5%
UTexas_run1	P/S	73.2%*
Yamraj_run1		70.7%
asjai_run5	S	69.8%
haLF_run2	S	69.4%*
RTM-DCU_run1		67.2%*
UANLPCourse_run2	S	48.7%

8.9. Participating systems: quantitative analysis (Relatedness)

ID	Composition	Accuracy (%)
Illinois-LH_run1	P/S	84.6%
ECNU_run1	S	83.6%
UNAL-NLP_run1		83.1%
SemantiKLUE_run1		82.3%
The_Meaning_Factory_run1	S	81.6%*
CECL_ALL_run1		80.0%
BUAP_run1	P	79.7%
UoW_run1		78.5%
Uedinburgh_run1	S	77.1%
UIO-Lien_run1		77.0%
FBK-TR_run3	P	75.4%
StanfordNLP_run5	S	74.5%
UTexas_run1	P/S	73.2%*
Yamraj_run1		70.7%
asjai_run5	S	69.8%
haLF_run2	S	69.4%*
RTM-DCU_run1		67.2%*
UANLPCourse_run2	S	48.7%

8.10. Qualitative analysis: balanced dataset (Entailment)

	Accuracy (%)		
ID	Full Dataset	Balanced Dataset	Variation
RTM-DCU_run1	67.2	70.4	+3.2
asjai_run5	69.8	72.8	+3.0
UTexas_run1	73.2	76.1	+2.9
UIO-Lien_run1	77.0	78.3	+1.3
Illinois_*compositional*_run ♣	**65.0**	**65.6**	**+0.6**
The_Meaning_Factory_run1	81.6	81.3	-0.3
Uedinburgh_run1	77.1	76.5	-0.6
Yamraj_run1	70.7	69.7	-1.0
UANLPCourse_run2	48.7	47.3	-1.4
StanfordNLP_run5	74.5	72.8	-1.7
UNAL-NLP_run1	83.1	81.0	-2.1
ECNU_*compositional*_run ■	**72.9**	**70.6**	**-2.3**
FBK-TR_run3	75.4	73.0	-2.4
UoW_run1	78.5	76.0	-2.5
SemantiKLUE_run1	82.3	79.7	-2.6
BUAP_run1	79.7	77.0	-2.7
ECNU_run1 ■	83.6	80.8	-2.8
haLF_run2	69.4	66.0	-3.4
Illinois-LH_run1 ♣	84.6	79.5	-5.1
CECL_ALL_run1	80.0	74.7	-5.3

8.11. Qualitative analysis: balanced dataset (Relatedness)

ID	Full Dataset	Balanced Dataset	Variation
asjai_run5	0.479	0.473	-0.006
Yamraj_run1	0.535	0.515	-0.020
The_Meaning_Factory_*compositional*_run ★	**0.608**	**0.583**	**-0.025**
RTM-DCU_run1	0.764	0.734	-0.030
UANLPCourse_run2	0.693	0.658	-0.035
StanfordNLP_run5	0.827	0.787	-0.040
ASAP_run1	0.628	0.586	-0.042
The_Meaning_Factory_run1 ★	0.827	0.783	-0.044
ECNU_*compositional*_run ■	**0.754**	**0.701**	**-0.053**
UTexas_run1	0.714	0.660	-0.054
UQeResearch_run1	0.642	0.585	-0.057
Illinois_*compositional*_run ♣	**0.463**	**0.397**	**-0.066**
CECL_ALL_run1	0.78	0.711	-0.069
SemantiKLUE_run1	0.78	0.711	-0.069
ECNU_run1 ■	0.828	0.758	-0.070
UNAL-NLP_run1	0.804	0.734	-0.070
BUAP_run1	0.697	0.625	-0.072
FBK-TR_run3	0.709	0.633	-0.076
Illinois-LH_run1 ♣	0.799	0.719	-0.080
UoW_run1	0.711	0.618	-0.093

8.12. Qualitative analysis: common errors (Entailment)

Table 18 Examples of the most difficult pairs in the Entailment Task.

A: A man is talking to a woman
B: A man and a woman are speaking
A: A black dog and a tan dog are fighting
B: Two dogs are fighting
A: Some women are dancing and singing
B: A woman is dancing and singing with other women
A: Two children and an adult are standing next to a tree limb
B: Three people are standing next to a tree limb
A: A man and a woman are sitting comfortably on the bench
B: Two people are sitting comfortable on the bench
A: A man and two women in a darkened room are sitting at a table with candle
B: The group of people is sitting in a room which is dim.
A: A basketball player is on the court floor and the ball is being grabbed by another one
B: Two basketball players are scrambling for the ball on the court

8.13. Qualitative analysis: common errors (Relatedness)

	Sentence A	Sentence B
Rel score $x \leq 2$	A man is playing baseball with a flute	A man is playing soccer
	A cat is looking at a store counter	A dog is looking around.
	Broccoli are being cut by a woman	A man is cutting tomatoes
	There is no man playing a game on the grass	A man is playing the guitar
Rel score $2 < x < 4.5$	The woman is penciling on eyeshadow	A woman is putting cosmetics on her eyelid
	A dog is chasing a ball in the grass	A dog with a ball is being chased in the grass
	A man is breaking a wooden hand against a board	A man is breaking wooden boards with his hand
	The man is riding a horse	A horse is riding over a man
Rel score $4.5 \leq x$	A man is riding on one wheel on a motorcycle	A person is performing tricks on a motorcycle
	The man is using a sledgehammer to break a concrete block that is on another man	A man is breaking a slab of concrete with a sledge hammer
	Many people are skating in an ice park	An ice skating rink placed outdoors is full of people

9. Admin

- Project presentations: Carlo suggests the 12th of May
- Written exam: the 17th at 10:30-12:30?