~~A Critical Review of~~ Natural Language Inference ~~Datasets~~ in SICK

Valeria de Paiva,
Nuance NAIL Lab, Sunnyvale
May, 2018

In May2016 we learned about Google's Parsey McParseface, the world's most accurate parser
94% accuracy vs. humans only 96-97%
Few years before: Manning on POS-tagging, accuracy at 97% (2011)
Manning/Nivre/Zeman and big cast on UDs
Also SICK (Marelli et al 2014):
simple corpus for compositional semantics…

Can we do this?

Natural Language Inference: our way

Valeria de Paiva, Nuance NAIL Lab, Sunnyvale
joint work with A. Kalouli (Konstanz, Germany), L. Real, (USP, Brazil)
May, 2018

Goals

- Separate logic knowledge, linguistic knowledge and world knowledge (as needed)
- Establish clear milestones on language and rules
- Improve lexical and semantic resources on a easy corpus
- Learn how to use embeddings

Natural Language Inference:

- Easier to detect inference between sentences than to decide on “good” semantic representations
- Need large, **high-quality,** corpora annotated for inference relations: entailment, contradiction, neutrality
- Fracas, SICK, SNLI, multiSNLI, SciTail,…
- **Can we trust these? Are they high-quality enough to learn from?**

NUANCE

Four papers investigating SICK

- Can we trust the corpus?
- What people consider logical inferences?
- Which kinds of inference are in the corpus?
- How can we calculate these kinds of inferences, if NOT using NNs?
- Which kinds of inference can we do using open source lexical resources like Wordnet, SUMO, JIGSAW, etc?

Outline

- Motivation
- The corpus SICK and its construction
- Processing SICK:
  - Simple entailments
  - Contradictions
  - One-word apart
- Analysis
- What's Next?

The Corpus SICK

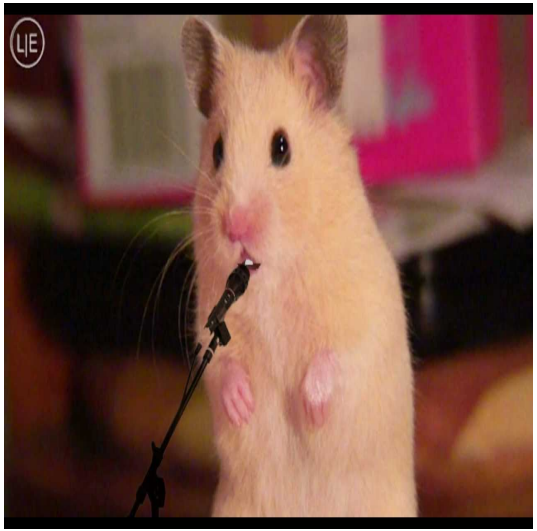
- SICK stands for Sentences Involved in Compositional Knowledge, result of 5-year European project COMPOSES(2011-2016)
- SICK data set consists of about 10,000 English sentence pairs, from captions of images (Flickr) and videos (youtube)
- Original sentences were **normalized** to remove unwanted linguistic phenomena: then sentences were **expanded** to obtain three new ones suitable for evaluation; all the sentences were paired to obtain the final data set.
- Each sentence pair was annotated for (relatedness and) entailment by means of (usual) crowdsourcing techniques.
- Entailment annotation led to 5595 *neutral* pairs, 1424 *contradiction* pairs, and 2821 *entailment* pairs
- Corpus and annotations available from http://clic.cimec.unitn.it/composes/sick.html

NUANCE

Why SICK?

We're interested in NL Inference

- SICK is a corpus marking the semantic relations we want: entailment, contradiction, neutrality
- The corpus is **simplified** to deal with these relations
- **Humans** were paid to do the man-in-the-street marking of these inference relations
- We want to use SICK as our baseline, but we need to **trust** this baseline
- Do revisions to SICK to use it as a baseline
- We assumed SICK was about commonsense, because of Flickr, but YouTube….



NUANCE

Creating the corpus SICK

- SICK sentences were “normalized” from captions and then
- “expanded” from a core set of sentences in visual corpora, via human constructed transformations such as adding passive or active voice, adding negations, adding adjectival modifiers, etc.
- Idea was to **simplify** the linguistic structure, and to **create** comparisons of different linguistic phenomena.
- They say: *[caption corpora] contain sentences (as opposed to paragraphs) that describe the same picture or video and are thus* **near paraphrases**, *are lean on named entities and rich in generic terms*.
- We were told the normalization/expansion was done by humans.
- We processed the corpus to check how well off-the-shelf tools would do in this corpus…

NUANCE

The SICK corpus

- Simple, short, common sense sentences: *People are walking* or *One man in a big city is holding up a sign and begging for money.*

- After de-duplication of sentences we have 6076 sentences, with 477 unique verb lemmas, 290 unique adjectives, 143 unique adverbs and 1100 unique nouns. (Processing with Stanford CoreNLP)

- Created from captions of pictures, SICK is about *daily* activities and entities; ones you can see in pictures taken by real people and which require **common sense** concepts: people, cats, dogs, running, climbing, eating, begging, etc.

(PWN should have all the lemmas? commonsense ontology should have all these concepts?)

sentences	words
4	3
129	4
257	5
580	6
821	7
847	8
622	9
2816	10 and more

NUANCE

Methodology: SICK corpus

- concrete, common sense, noncomplex sentences
  near 10K pairs
  benchmark for compositional Distributional Semantics
  (few complex linguistics, no named entities, no temporal issues, no MWEs,…)

- Pairs annotated for *similarity* degree and inference relations *(e,c,n)*

A = Kids in red shirts are playing. *B = Children in red shirts are playing*	*AeBBeA*
A = A man isn't sitting at the table. *B = A man is sitting at a table.*	*AcBBcA*
A= A silver airplane is landing. *B= A plane is landing.*	*AeBBnA*

NUANCE

Processing the sentences
with Stanford UDs

- Full details, code and data, can be found in our GitHub repository https://github.com/kkalouli/SICK-processing
- We run the sentences through Stanford's CoreNLP and use Enhanced Dependencies++, https://nlp.stanford.edu/~sebschu/pubs/schuster-manning-lrec2016.pdf.
- From the CoNLL representation of each sentence we obtain its bag-of-linguistic concepts using Princeton WordNet (PWN)
- We use Word Sense Disambiguation (WSD) via JIGSAW to pick a favorite sense, https://github.com/pippokill/JIGSAW.
- We intend to use PWN to SUMO mappings to obtain logical concepts some time soon…

Previous processing outputs (Parsey McParseface/Freeling/UDpipe) also available and can be used for comparisons

Preprocessing+UDparsing

(Tokenization is not canonical)

	Parsey +FreeL	CoreNLP+ Jigsaw
Verb lemmas	513	477
Noun lemmas	1076	1100
Adjective lemmas	270	290
Adverb lemmas	149	143

Construction of SICK

How did they say they construct their corpus? (Marelli et al LREC2014 paper)

Rule	Example
Replace possessive pronouns with the word they stand for or with a determiner.	S0: "*A man is standing outside his house*" S1: "*A man is standing outside the house*"
Replace Named Entities with a word that stands for the class.	S0: "*A woman is playing Mozart*" S1: "*A woman is playing classical music*"
In order to avoid generic sentences , transform all non-stative verb tenses into present continuous.	S0: "*Birds land on clothes lines*" S1: "*Birds are landing on clothes lines*"
Replace complex verb constructions into simpler ones.	S0: "*A man is attempting to surf down a hill made of sand*" S1: "*A man is surfing down a hill made of sand*"
Simplify verb phrases with modals and auxiliaries.	S0: "*A kid has to eat a vegetable soup*" S1: "*A kid is eating a vegetable soup*"

Construction of SICK

How did they construct their corpus? More “normalization” rules

Replace phrasal verbs with a synonym if verb and preposition are not adjacent.	S0: “*A man is sorting the documents out*” S1: “*A man is organizing the documents*”
Remove multiword expressions.	S0: “*A person is playing guitar right now*” S1: “*A person is playing guitar*”
Remove dates and numbers; if the number is a determiner write it in letters.	S0: “*3 people are on a small boat enjoying the view*” S1: “*Three people are on a small boat enjoying the view*”
Turn subordinates into coordinates.	S0: “*A faucet is running while a bird is standing in the sink below*” S1: “*A faucet is running and a bird is standing in the sink below*”
Turn non-sentential descriptions into sentences.	S0: “*An airplane in the air*” S1: “*An airplane is flying in the air*”
Remove indirect interrogative and parenthetical phrases.	We did not find any instance in the data sets

SICK normalization rules

1. *A man is standing outside **his** house* ➔ *A man is standing outside **the house***
   (still 322 occurrences of nmod:poss plus 41 genitives "*somebody's legs*", how bad?)

2. *A woman is playing Mozart* ➔ *A woman is playing classical music*

(still missing Seadoo, bmx, ATVs, Canon [camera]…) also **no** *biker, motocross, wheelie, corndog, wetsuit, jetski, kiddie, footbag, kickboxing* in PWN

1. *Birds land on clothes lines* ➔ *Birds are landing on clothes lines*
   (**bad idea**, does not avoid generic sentences. not totally done either)

2. *A man is **attempting to** surf down a hill made of sand* ➔
   A man is surfing down a hill made of sand
   (definition of **complex verb construction**? 90 xcomps)

3. *A kid **has** to eat a vegetable soup* ➔*A kid is eating a vegetable soup (still 4 modals)*

NUANCE

SICK expansion rules

Original pair	
S0a: *A sea turtle is hunting for fish*	**S0b:** *The turtle followed the fish*
Normalized pair	
S1a: *A sea turtle is hunting for fish*	**S1b:** *The turtle is following the fish*
Expanded pair	
S2a: *A sea turtle is hunting for food*	**S2b:** *The turtle is following the red fish*
S3a: *A sea turtle is not hunting for fish*	**S3b:** *The turtle isn't following the fish*
S4a: *A fish is hunting for a turtle in the sea*	**S4b:** *The fish is following the turtle*

Table 1: Example of output of data set creation process.

The theory of SICK

How well did they do the job of simplifying the language?

- The curators of the corpus also made an effort to reduce the amount of “encyclopedic knowledge” about the world that is needed to do inference.
- They say “*To ensure the quality of the data set, all the sentences were checked for grammatical or lexical mistakes and disfluencies by a native English speaker*.”

- Reasonably well, we expect? **Not really!!**
- Many sentences do not make sense. Many are not commonsense at all.

- But the transformations adopted are also very debatable.

“The monkey is brushing the dog”

NUANCE

The theory of SICK 1

Meaning-preserving Transformations (from their LREC2014 paper)

Turn active sentences into passive sentences and viceversa.	S1: "*A man is driving a car* " S2: "*The car is being driven by a man*"
Replace words with near synonyms or similar words.	S1: "*A young boy is jumping into water* " S2: "*A young kid is jumping into water*" S1: *A man and two women in a darkened room are sitting at a table with candles* S2: *A man and two women in a dark room are sitting at a table with candles*
Add modifiers that do not radically alter the meaning of the sentence.	S1: "*A deer is jumping a fence* " S2: "*A wild deer is jumping a fence*" S1: "*A woman is tapping her fingers*" S2: "*A woman is tapping her fingers nervously*"
Expand agentive nouns.	S1: "*A soccer player is kicking a ball into the goal*" S2: "*A person who plays soccer is kicking a ball into the goal*"
Turn compounds into relative clauses.	S1: "*A woman is using a sewing machine*" S2: "*A woman is using a machine made for sewing*"
Turn adjectives into relative clauses.	S1: "*Two men are taking a break from a trip on a snowy road* " S2: "*Two men are taking a break from a trip on a road covered by snow*"
Replace quantifiers with others that have a similar meaning.	S1: "*The surfer is riding a big wave*" S2: "*A surfer is riding a big wave*"

NUANCE

The theory of SICK 2

Meaning-altering Transformations

Change determiners with their opposite. $\{the, a, all, every, some, a\ few\} \Rightarrow \{no\}$, $\{no\} \Rightarrow \{every, each\}$, $\{many\} \Leftrightarrow \{few\}$, $\{much\} \Leftrightarrow \{little\}$.	S1: "*A dog is walking along a snowdrift*" S3: "*There is no dog walking along a snowdrift*"
Replace words with semantic opposites.	S1: "*The girl is spraying the plants with water*" S3: "*The boy is spraying the plants with water*" S1: "*A plane is taking off*" S3: "*A plane is landing*"
Scramble words: switch the arguments of a transitive verb, switch and mix modifiers, exploit verb transitive/intransitive alternations, exploit	S1: "*The turtle is following the fish*" S4: "*The fish is following the turtle*" S1: "*A man with a jersey is dunking the ball at a basketball game*" S4: "*The game of basketball consists of a ball being dunked by a man*

Table from LREC2014 paper

NUANCE

The SICK normalization rules

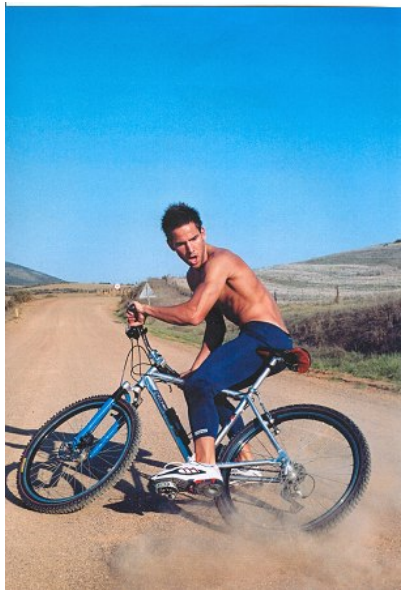
1. *A man is sorting the documents out* ➔ *A man is organizing the documents*
   (PWN not very useful. Sometimes have verb+prep, some times not. How important? Other resources?)
2. *A person is playing guitar right now* ➔ *A person is playing guitar*
   (which mwes? still 1.2K compounds, 432 unique, 84 mwes in PWN)
3. (no dates in corpus, still several numbers. useful?)

4. *A faucet is running while a bird is standing in the sink below* ➔*A faucet is running and a bird is standing in the sink below*
   (ok, done, no sconj in corpus. useful?)
5. *An airplane in the air* ➔*An airplane is flying in the air*
   *(*still some captions e.g *A couple standing on the curb*)
6. they say they didn't find either parenthetical phrases or indirect interrogatives, but e.g. *Four middle eastern children, three girls and one boy, are climbing on the grotto with a pink interior.*

NUANCE

Theory vs Practice

- Passive and active voice should work, but UDs deal with passive badly
- “dark room” = “darkened room” but how many like it in PWN?
- Vacuous modifiers? Wild deer = deer, **world knowledge**!
- Expand agentives? Bad idea!
- Compounds into relative clauses? Bad idea
- Adjectives into relative clauses? Bad idea
- Similar quantifiers? Yes!
- Opposite determiners? ok
- “semantic opposites”? Hard to decide…
- Scramble? Terrible idea!



NUANCE

Analysis of SICK

There are many atrocious sentences that do not make sense at all:

A person is ignoring the motocross bike that is lying on its side and there is no one is racing by; (the third ``is" is a typo, but the rest?) or *There is no man wearing clothes that are covered with paint or is sitting outside in a busy area writing something.*

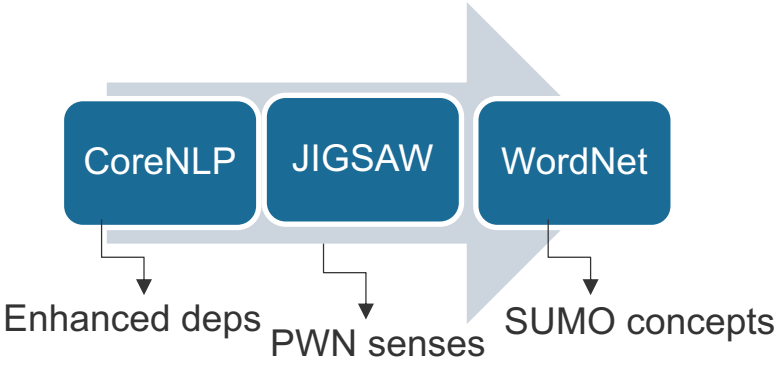
How can we measure how many bad sentences are there? We decided to investigate two sub-corpora.

First the single-sided **entailments** and secondly the **contradictions** that are logically incorrect.

SICK annotation & problems

1424 pairs of contradictions (*AcBBcA*)
1300 pairs of bi-entailment (*AeBBeA*)
1513 pairs of single entailment(*AeBBnA*)
4992 pairs of neutrals (*AnBBnA)*

Annotation Problems:
Noisy data, lack of strict guidelines, genuine semantic differences, etc
→ '**asymmetric**' contradictions
611 pairs are **not logical**:
(30% of all contradictions!)
flutes are not guitars!



NUANCE

Single Entailments IWCS2017

- Manually checked all single entailments 1513 AeBBnA pairs
- Taxonomy of errors:
- E.g. Non-binding referents:
  *A = **An** Asian woman in **a** crowd is not carrying a black bag*
  *B = **An** Asian woman in **a** crowd is carrying a black bag. AcBBnA*
- Definitions, e.g.
  *A = There is no man on a bicycle riding on the **beach**.*
  *B = A person is riding a bicycle in the **sand beside the ocean**. AeBBcA*
- "Privative" adjectives and nouns: contradict the noun they're modifying,
  *A = A **cartoon airplane** is landing*
  B = A plane is landing. AeBBnA
- **Noisy** data
- **12% pairs** needed correction (corrected ones are in GitHub)

Single Entailments ICWS2017

Contradictions CONLI2017

PWN Easy Inferences LREC2018

GKR for SICK NLCS2018

NUANCE

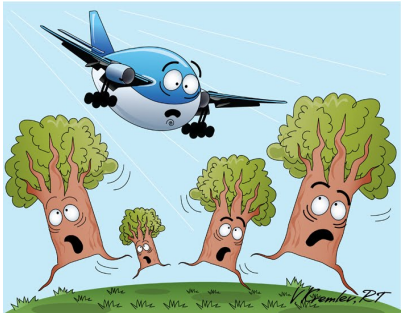
Taxonomy of problems

1. Non-binding referents: no given referent or context to judge the pairs, e.g. ***An*** *Asian woman in* ***a*** *crowd is not carrying a black bag.* ***vs***. ***An*** *Asian woman in* ***a*** *crowd is carrying a black bag.* AcBBnA
2. "Alternative" concepts, e.g.*The lady is cracking an egg into a* ***bowl****.* ***vs****.The lady is cracking an egg into a* ***dish****.* AeBBcA
3. Issues with some definitions, e.g. *There is no man on a bicycle riding on the* ***beach****. vs. A person is riding a bicycle in the* ***sand beside the ocean****.* AeBBcA
4. Plain errors, e.g. *The blond girl is dancing behind the sound equipment.* vs. *The blond girl is dancing in front of the sound equipment.* AcBBnA
5. Ungrammatical sentences, e.g.*The black and white dog isn't running and there is no person standing behind*
6. Nonsensical sentences, e.g.: *A motorcycle is riding standing up on the seat of the vehicle*

NUANCE

More Issues

- compound nouns: deverbal adjectives modifying nouns, e.g.*The microphone in front of the **talking parrot** is being muted. B = A parrot is speaking.* AeBBnA
- privative [Partee] adjectives/nouns contradicting their modifying noun, *A **cartoon airplane** is landing. **vs** A plane is landing* AeBBnA
- quantifier scope, e.g *Two bikes are being ridden by two people. vs.Two people are riding a bike.*
- cultural definitions, e.g *Different teams are playing **football** on the field. vs Two teams are playing **soccer**.* AeBBnA
- agentive nouns, e.g. *cyclist* and *model* : Everyone who rides a bicycle is a cyclist, but is everyone who poses for a photo a model?
- even the simple SICK pairs need more than lexical semantics, e.g. *One man is turning on the microwave. **vs**. The buttons of a microwave are being pushed by a man.*



Previous work: Contradictions

- Detecting conflicting or contradictory statements is a fundamental text understanding task within many applications (Condoravdi et al.,2003)
- Contradictions in logic are symmetric: if proposition A is contradictory to B, then B must be contradictory to A
- *Two children are lying in the snow and are making snow angels*. **vs** *There is no child lying in the snow and making snow angels*.
- *A man isn't sitting comfortably at the table*. **vs.** *A man is sitting comfortably at a table.*
- 1424 pairs are AcBBcA

NUANCE

Contradictions 2

- 611 pairs are **asymmetric** contradictions: what?
- out of 9840 may seem few (around 6%),
- but around 30% of all contradictions found
- Analysis - taxonomy
- Different procedures to not lose many pairs/labels

NUANCE

Contradictions CONLI 2017

- Contradictions are hard
- MUST: Associate referents in both sentences.
- Assume pairs are talking about same event and entities, no matter whether definite or indefinite markers (the, a) are used
- Only find contradictions in sentences that are '**close enough**'.
- BUT: difficult to define close enough, predicates 'contradictory in context' **need commonsense**
- Punt on non-atomic sentences
- Re-annotated at least 611 pairs

Single Entailments ICWS2017

Contradictions CONLI2017

PWN Easy Inferences LREC2018

GKR for SICK NLCS2018

NUANCE

WordNet “Easy” Inferences (LREC2018)

- How easy is “Easy”?
- Corrected sub-corpus of 2936 pairs of “one-word apart” pairs
- 30% original corpus
- Turkers don’t want to work more than necessary
- 1,6K pairs checked by heuristics+PWN
- 1,4K pairs relationship determined by words apart, but PWN+heuristics don’t know what it is.
- Mostly “synonymy/antonymy-in-context”
- But also meronym (not in system) , prepositions, compounds, many long tail phenomena…
- Suggest few improvements to WordNet (e.g. rope~cord, shoot~ fire)

NUANCE

GKR for SICK

- Work on SICK conceived as a trusted baseline for work on GKR
- Quite a bit to develop. how are we doing?
- Comparing AMR, ProPs and GKR
- Conjunction, disjunction and negation in this installment
- Wait for JULY!

Conclusions?

- Want to trust our baseline. Need golden standards that can be trusted.
- Contradictions ought to be symmetric.
- Corpus design: the explicitation of the referents of a sentence plays a huge role, especially when dealing with contradictions
- Corpus annotation: must have controlling mechanisms and guidelines
- Which kind of contradiction does a system need?

NUANCE

Conclusions

Semi-automatic investigation of SICK

- want to make corpus a **real golde**n standard, not there, yet…
- conceptualize the limits of **lexical semantics**
- understand better challenges of NLI

- All corpora suffer from noisy annotations:
- Data curation efforts are essential to establish trustworthy baselines;
- Cleaning up data ensures that corrected mistakes can be used as guidelines for future corpora.
- One-word-apart is useful method, want to check other NLI corpora

Next Steps:

- **Summer Internship**
- logic+vector embeddings for inference

NUANCE

CODA: No NNs for NLI?

1. https://arxiv.org/abs/1805.02266. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences (may2018) (Goldberg)
2. https://arxiv.org/abs/1805.01042 Hypothesis Only Baselines in Natural Language Inference (may) JHU hypothesis-only model, subsets of SNLI
3. https://arxiv.org/abs/1804.08117 Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment (apr)

4. https://arxiv.org/abs/1803.02324 Annotation Artifacts in Natural Language Inference Data (apr 16) JHU+Bowman, use "hard pairs" for inference (ones that do not get predicted by hypothesis only)
5. https://arxiv.org/pdf/1802.04302.pdf (Goodman) (feb 12) new corpus, comparatives, corpus at https://github.com/ishita-dg/ScrambleTestt
6. https://arxiv.org/pdf/1803.05355.pdf FEVER

Thank you

SR principles

1 .Multilingual is essential -- SR need to be satisfactory (linguistic analysis grounds) for individual languages.
2. Parallelism is good -- SR need to provide a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
3. Easiness is essential – SR must be suitable for rapid, consistent annotation by a incompetent human annotator.
4. Efficiency is essential -- SR must be suitable for computer parsing with high accuracy.
5. Easiness of understanding-- SR must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing
6. Applications must be supported-- SR must support well downstream language understanding tasks (relation extraction, reading comprehension, machine translation,etc)

References

- M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli. **A SICK cure for the evaluation of compositional distributional semantic models**. *LREC 2014*, 2014.
- I. Niles and A. Pease. **Toward a Standard Upper Ontology**. In Welty and Smith, editors, *FOIS-2001*, pages 2–9, 2001.
- P. Basile, M. de Gemmis, A. Gentile, P. Lops, and G. Semeraro. **Uniba: JIGSAW algorithm for word sense disambiguation** *SemEval-2007*, Prague, Czech Rep.
- S. Schuster and C. D. Manning. **Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks**. *LREC2016.*
- Aikaterini-Lida Kalouli, Livy Real, Valeria de Paiva. **Textual Inference: getting logic from humans**. IWCS, September 2017.
- Aikaterini-Lida Kalouli, Livy Real, Valeria de Paiva. **Correcting Contradictions**. CONLI Workshop, 2017. *Montpellier, France.*
- Aikaterini-Lida Kalouli, Livy Real, Valeria de Paiva. **Annotating Logic Inference Pitfalls** Workshop on Data Provenance/Annotation in Computational Linguistics 2018,
- Aikaterini-Lida Kalouli, Livy Real, Valeria de Paiva. **WordNet for "Easy" Textual Inferences,** Globalex Workshop, LREC2018
- Aikaterini-Lida Kalouli, Dick Crouch, Valeria de Paiva, Livy Real. **Graphical Knowledge Representations for SICK**, NLCS 2018, to appear

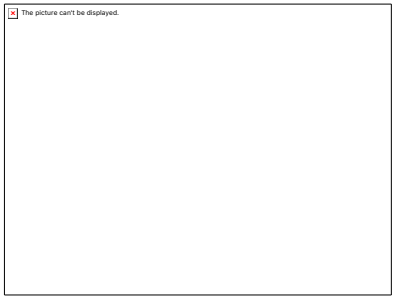
NUANCE

Composing entailments

- We restrict ourselves, to begin with, to pairs of sentences that differ by a single word.
- Even this is not enough as can be seen in the example: *A man is passionately playing guitar* compared to *A man is playing acoustic guitar*.
- Clearly an *acoustic guitar* is a kind of *guitar*
- Clearly *X is passionately playing guitar* implies *X is playing guitar*;
- But *X is passionately playing guitar* does not imply *X is playing acoustic guitar*.
- Neither *X is playing acoustic guitar* implies *X is passionately playing guitar* . The sentences are neutral wrt each other

3-4-5: a hand-checked sub-corpus

390 sentences (385 nsubj+ 5 nsubjpass)



- Only one conjunction:

Paper and scissors both cut

- 13 copulas (4 of them are wrong:

The man is training, The man is rock climbing,
A woman is grating carrots, The woman is dicing garlic)

- 19 negations marked, but 26 neg expletives, 10 nobodies, 1 no person ➔ should be 56 negations, needs work, we know.
- Compounds? Have 20 nns, some are real compounds:
- *ping pong, golden retriever, sumo wrestlers, tiger cub, sumo ringers, baby pandas, cartoon airplane*
- Some are processing mistakes, gerund for the noun, lots (14 in 390) *hamster singing, lion walking, panda climbing, etc.*
- Also 9 cases of particle verbs (need to update PWN?)

Problems pale in comparison to WSD

NUANCE

