SemEval 2012
STS task

http://www.cs.york.ac.uk/semeval-2012/task6/

Eneko Agirre
Daniel Cer
Mona Diab
Bill Dolan

Dates

Trial dataset: 20 October
Call for participation: 25 October
Training dataset + test scripts: 31 December
Start of Evaluation period: 18 March
End of Evaluation period: 1 April
Paper due: 11 April

7-8 June *SEM conference (with NAACL)

Outline

- Description of the task
- Source Datasets
- Annotation
  - Instructions
  - Pilot
  - AMT
- Quality of annotation

Description of the task

- Given two sentences of text, s1 and s2:
  - Return a similarity score
  - ... and an optional confidence score
- Evaluation:
  - Correlation (Pearson)
    with average of human scores

Source Datasets

- We wanted to reuse already existing datasets
- Textual entailment

> T:The Christian Science Monitor named a US journalist kidnapped in Iraq as freelancer Jill Carroll.
> H:Jill Carroll was abducted in Iraq.

- Paraphrase: MSR paraphrase and video
- Machine translation evaluation: WMT

MSR paraphrase corpus

- Widely used to evaluate text similarity algorithms
- Gleaned over a period of 18 months from thousands of news sources on the web.
- 5801 pairs of sentences
  - 70% train, 30% test
  - 67% yes, %33 no
    - completely unrelated semantically, partially overlapping, to those that are almost-but-not-quite semantically equivalent.
  - IAA 82%-84%
- (Dolan et al. 2004)

MSR paraphrase corpus

- The Senate Select Committee on Intelligence is preparing a blistering report on prewar intelligence on Iraq.
- American intelligence leading up to the war on Iraq will be criticised by a powerful US Congressional committee due to report soon, officials said today.

- A strong geomagnetic storm was expected to hit Earth today with the potential to affect electrical grids and satellite communications.
- A strong geomagnetic storm is expected to hit Earth sometime %%DAY%% and could knock out electrical grids and satellite communications.

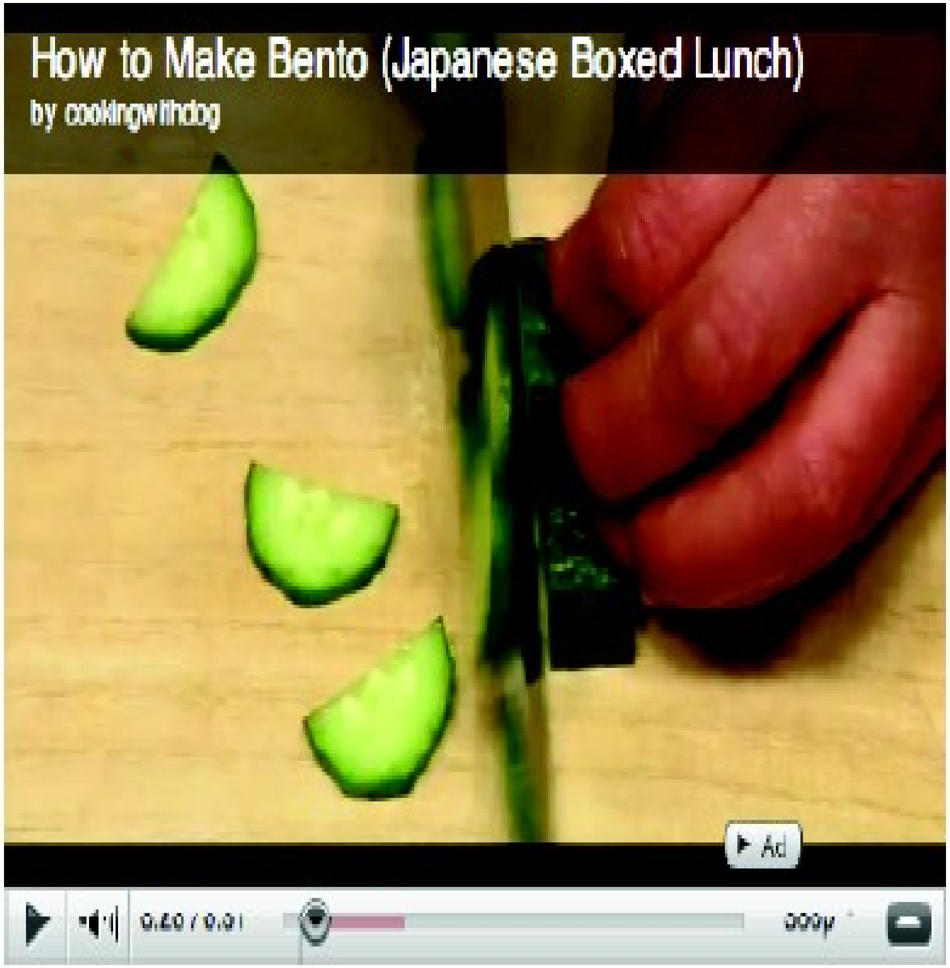
MSR paraphrase corpus

- Methodology:
  - Rank pairs according to string similarity
    - Algorithms for Approximate String Matching", E. Ukkonen, Information and Control Vol. 64, 1985, pp. 100-118.
  - Five bands (0.8 – 0.4 similarity)
  - Sample equal number of pairs from each band
  - Repeat for paraphrases / non-paraphrases
  - 50% from each
- 750 pairs for train, 750 pairs for test

MSR Video Description Corpus

- Show a segment of YouTube video
  - Ask for one-sentence description of the main action/event in the video (AMT)
  - 120K sentences, 2,000 videos
  - Roughly parallel descriptions (not only in English)
- (Chen and Dolan, 2011)

MSR Video Description Corpus



- A person is slicing a cucumber into pieces.
- A chef is slicing a vegetable.
- A person is slicing a cucumber.
- A woman is slicing vegetables.
- A woman is slicing a cucumber.
- A person is slicing cucumber with a knife.
- A person cuts up a piece of cucumber.
- A man is slicing cucumber.
- A man cutting zucchini.
- Someone is slicing fruit.

MSR Video Description Corpus

- Methodology:
  - All possible pairs from the same video
  - 1% of all possible pairs from different videos
  - Rank pairs according to string similarity
  - Four bands (0.8 – 0.5 similarity)
  - Sample equal number of pairs from each band
  - Repeat for same video / different video
  - 50% from each
- 750 pairs for train, 750 pairs for test

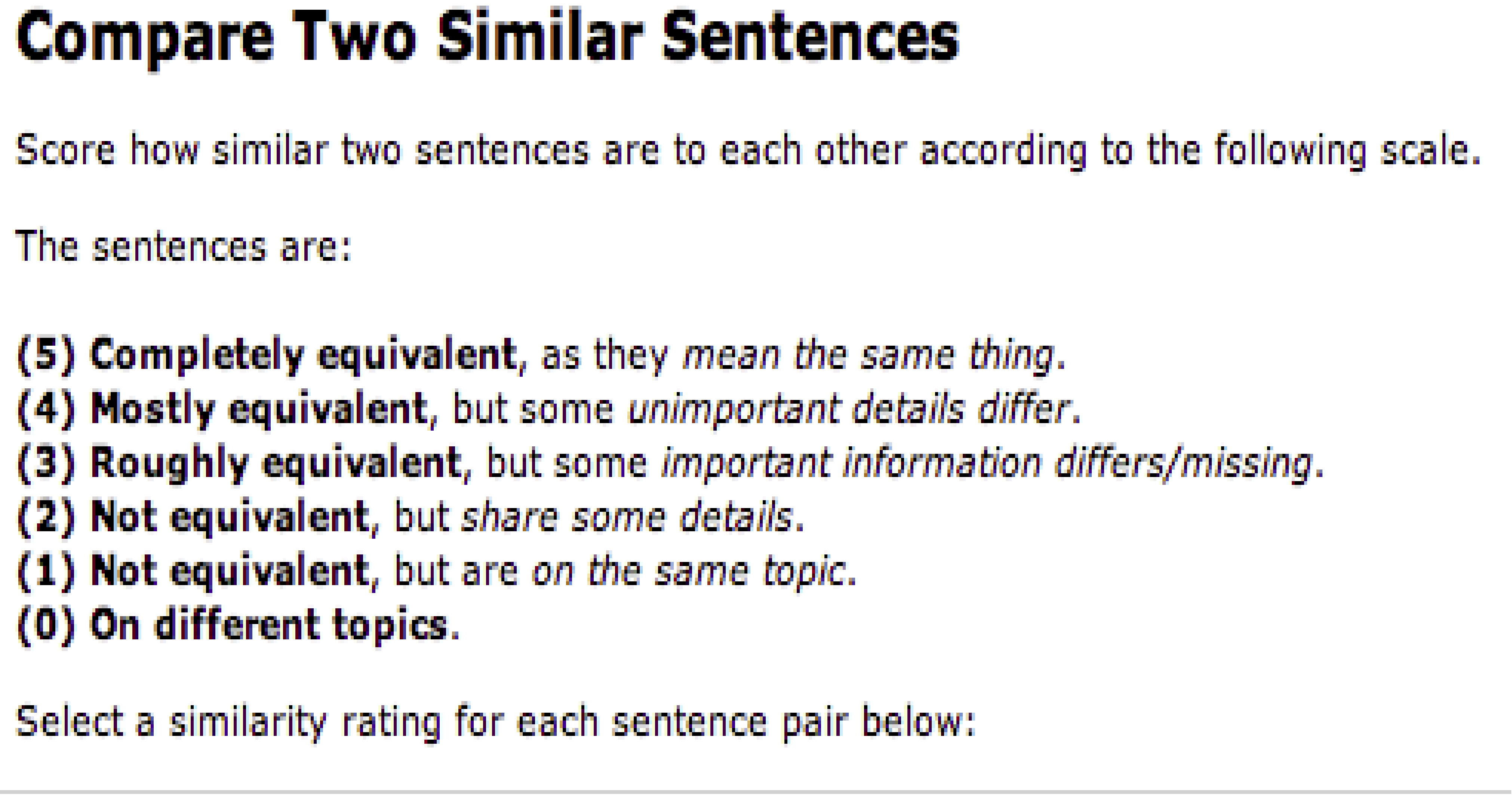
WMT: MT evaluation

- Pairs of segments (~ sentences) that had been part of the human evaluation for WMT systems
  - a reference translation
  - a machine translation submission
- To keep things consistent, we just used French to English system submissions translation
- Train contains pairs in WMT 2007
- Test contains pairs with less than 16 tokens from WMT 2008
- Train and test come from Europarl

WMT: MT evaluation

- The only instance in which no tax is levied is when the supplier is in a non-EU country and the recipient is in a Member State of the EU.
- The only case for which no tax is still perceived "is an example of supply in the European Community from a third country.
- Thank you very much, Commissioner.
- Thank you very much, Mr Commissioner.

Annotation



Compare Two Similar Sentences

Score how similar two sentences are to each other according to the following scale.

The sentences are:

(5) Completely equivalent, as they *mean the same thing*.
(4) Mostly equivalent, but some *unimportant details differ*.
(3) Roughly equivalent, but some *important information differs/missing*.
(2) Not equivalent, but *share some details*.
(1) Not equivalent, but are *on the same topic*.
(0) On different topics.

Select a similarity rating for each sentence pair below:

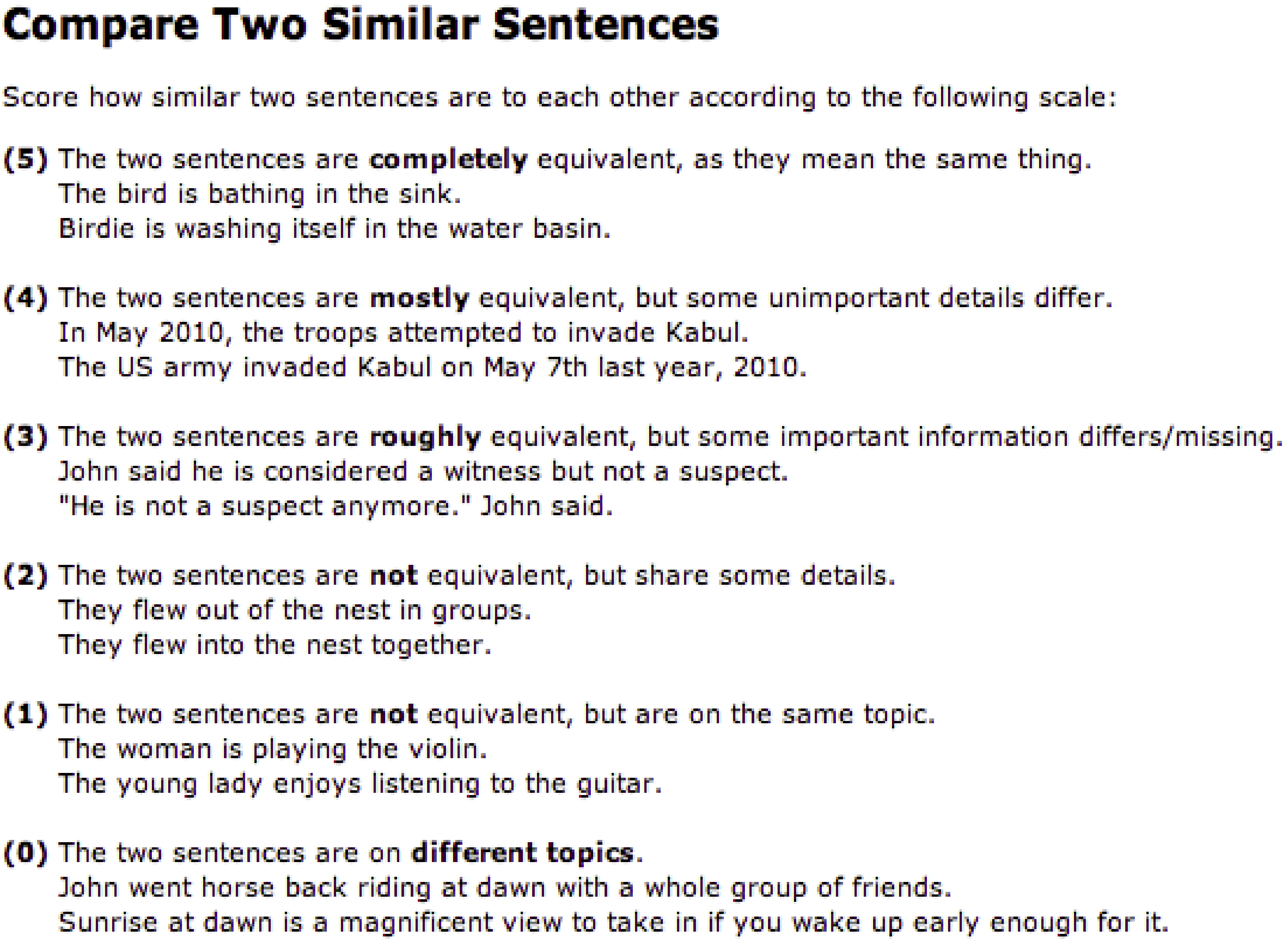
Pilot

- Mona, Dan, Eneko
- ~200 pairs from three datasets
- Pairwise agreement:

  ```
  • GS:dan     SYS:eneko     N:188 Pearson: 0.874
  • GS:dan     SYS:mona      N:174 Pearson: 0.845
  • GS:eneko   SYS:mona      N:184 Pearson: 0.863
  ```

- Agreement with average of rest of us:

  ```
  • GS:average  SYS:dan   N:188 Pearson: 0.885
  • GS:average  SYS:eneko N:198 Pearson: 0.889
  • GS:average  SYS:mona  N:184 Pearson: 0.875
  ```



Compare Two Similar Sentences

Score how similar two sentences are to each other according to the following scale:

(5) The two sentences are **completely** equivalent, as they mean the same thing.
The bird is bathing in the sink.
Birdie is washing itself in the water basin.

(4) The two sentences are **mostly** equivalent, but some unimportant details differ.
In May 2010, the troops attempted to invade Kabul.
The US army invaded Kabul on May 7th last year, 2010.

(3) The two sentences are **roughly** equivalent, but some important information differs/missing.
John said he is considered a witness but not a suspect.
"He is not a suspect anymore." John said.

(2) The two sentences are **not** equivalent, but share some details.
They flew out of the nest in groups.
They flew into the nest together.

(1) The two sentences are **not** equivalent, but are on the same topic.
The woman is playing the violin.
The young lady enjoys listening to the guitar.

(0) The two sentences are on **different topics**.
John went horse back riding at dawn with a whole group of friends.
Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.

Pilot with turkers

- Average turkers with our average:
  - N:197 Pearson: **0.959**
- Each of us with average of turkers:
  - dan N:187 Pearson: 0.937
  - eneko N:197 Pearson: 0.919
  - mona N:183 Pearson: 0.896

Working with AMT

- Requirements:
  - 95% approval rating for their other HITs on AMT.
  - To pass a qualification test with 80% accuracy.
    - 6 example pairs
    - answers were marked correct if they were within +1/-1 of our annotations
  - Targetting US, but used all origins
- HIT: 5 pairs of sentences, $ 0.20, 5 turkers per HIT
- 114.9 seconds per HIT on the most recent data we submitted.

Working with AMT

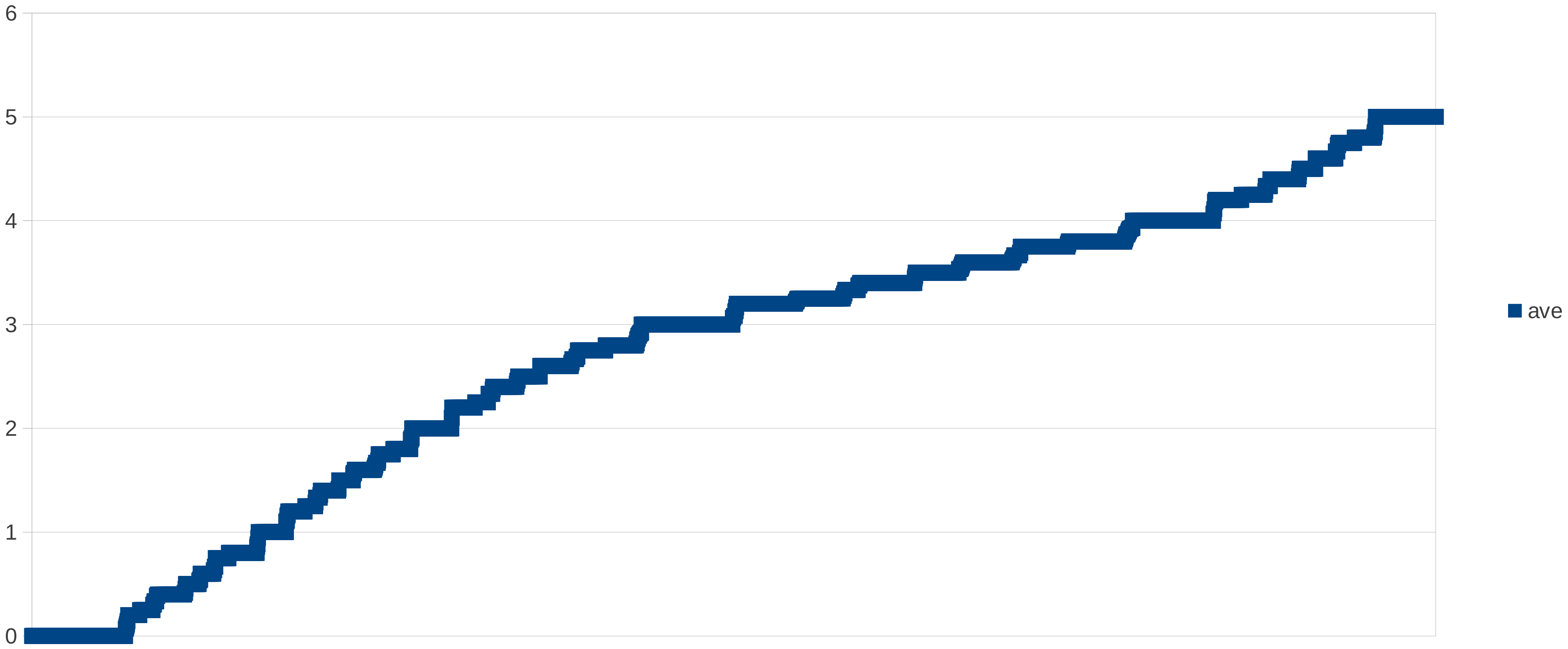
- Quality control
  - Each HIT contained one pair from our pilot
  - After the tagging we check correlation of individual turkers with our scores
  - Remove annotations of low correlation turkers
    - A2VJKPNDGBSUOK N:100 Pearson: -0.003
  - Later realized that we could use correlation with average of other Turkers

Assessing quality of annotation

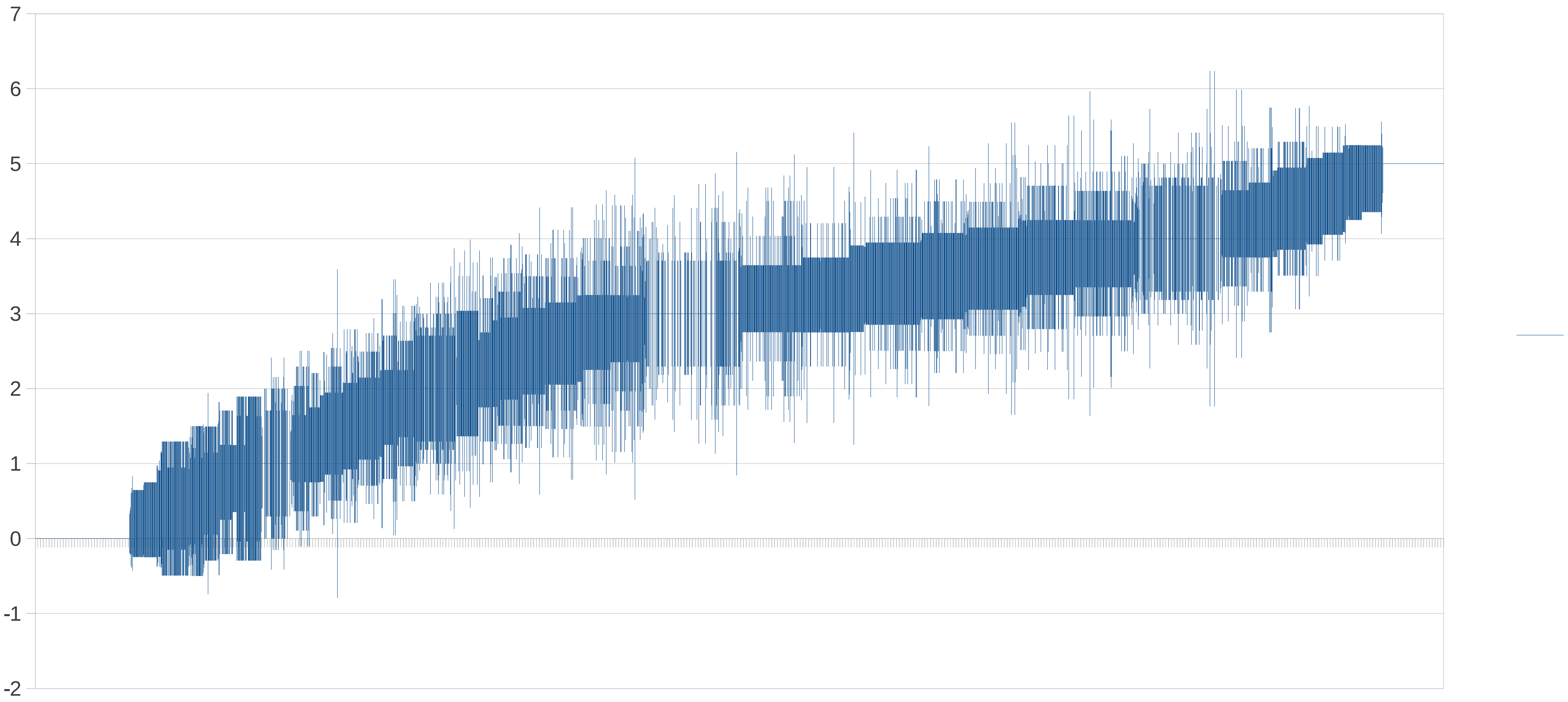
Assessing quality of annotation

- MSR datasets
  - Average 2.76
  - 0:2228
  - 1:1456
  - 2:1895
  - 3:4072
  - 4:3275
  - 5:2126

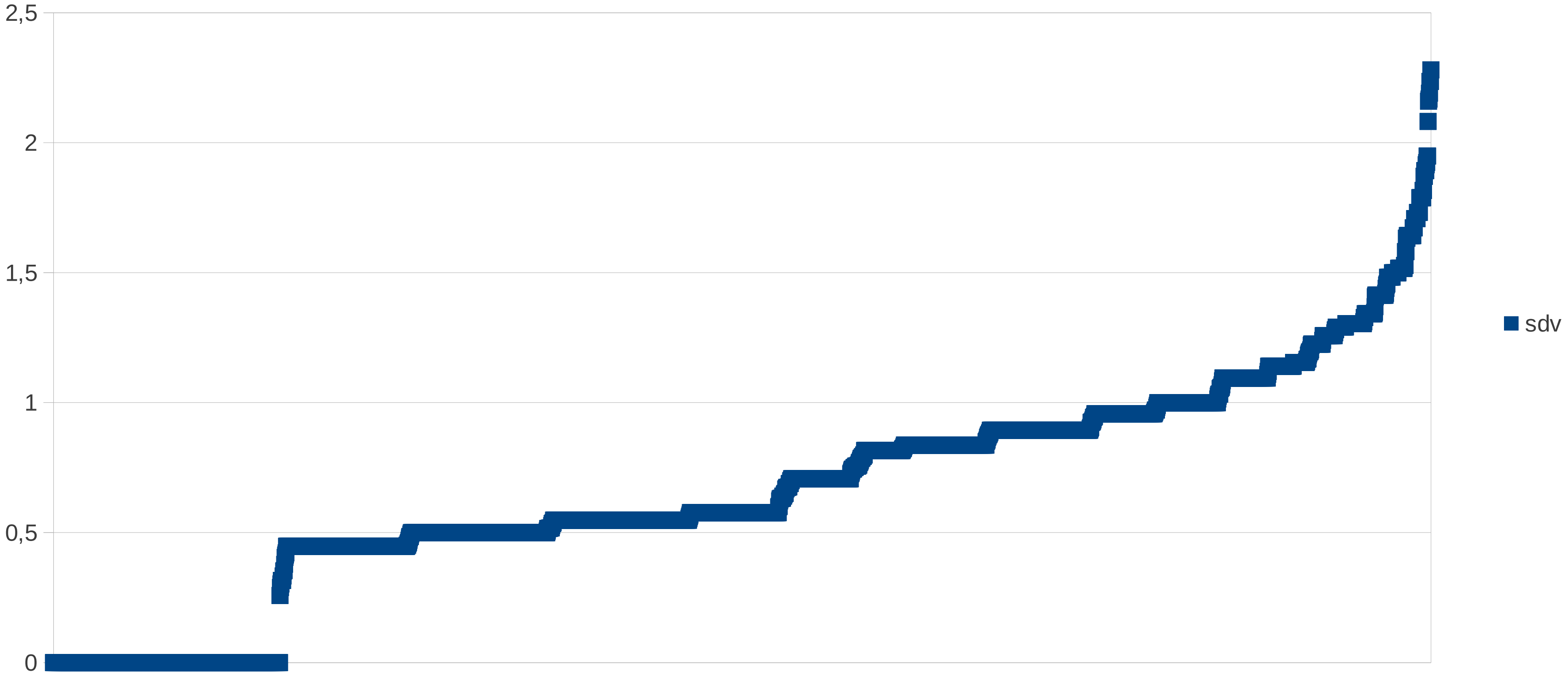
Average (MSR data)



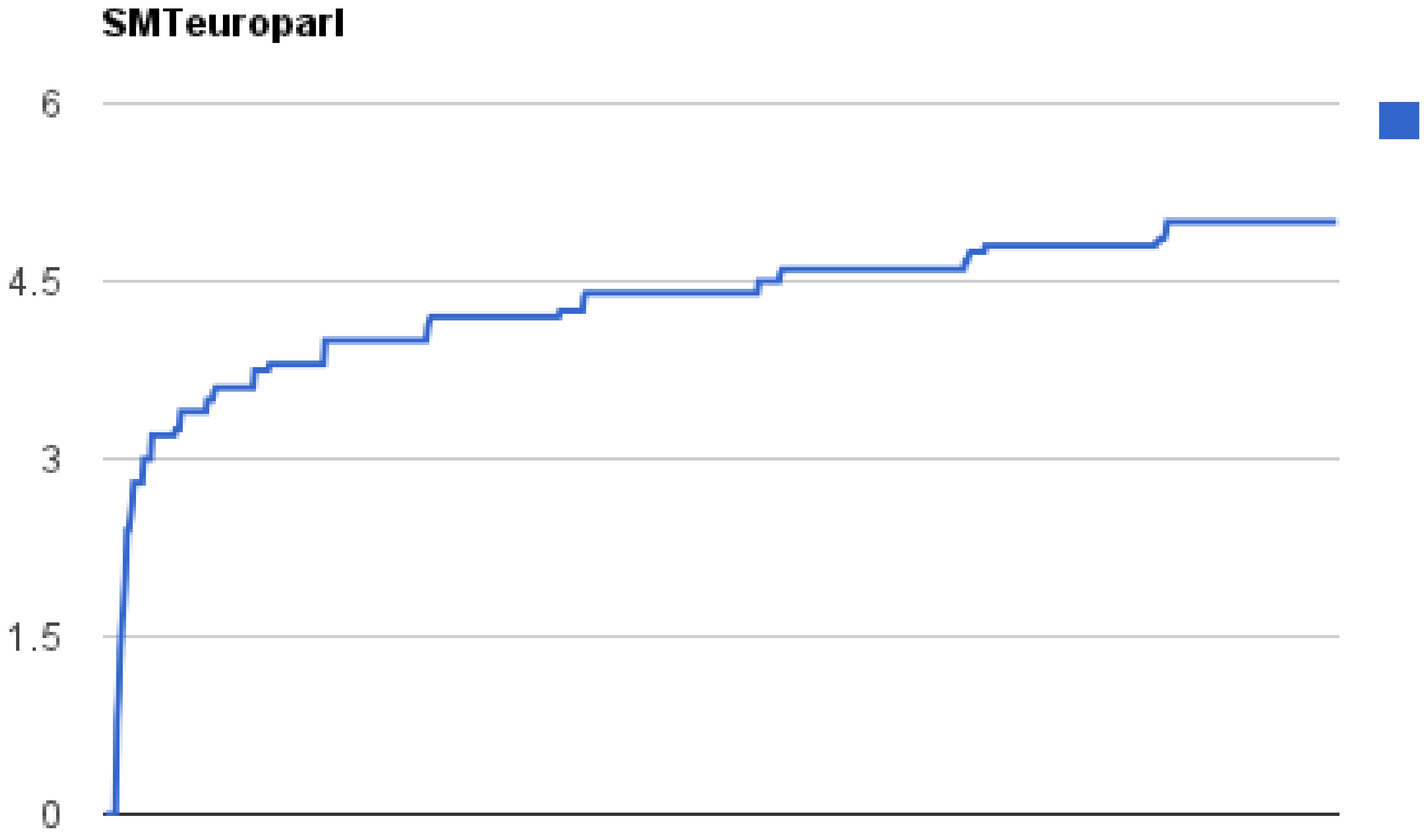
Standard deviation (MSR data)



Standard deviation (MSR data)



Average SMTeuroparl



Conclusions

- Wealth of annotated data:
  - 1500 pairs from MSRpar and MSRvid (each)
  - ca. 1000 pairs from WMT 2007/2008
  - Surprise datasets (ca. 1500 pairs)
- Current work:
  - Correlation with MSR paraphrase
  - Correlation with WMT
- Open issue:
  - Alternatives to the opportunistic method
  - How to collect pairs of sentences?
  - How to collect pairs of sentences related to a single phenomenon (e.g. Negation)?

SemEval 2012
STS task

http://www.cs.york.ac.uk/semeval-2012/task6/

Eneko Agirre
Daniel Cer
Mona Diab
Bill Dolan