Mining Wikipedia for Large-scale Repositories of Context-Sensitive Entailment Rules

Milen Kouylekov[1], Yashar Mehdad[1;2], Matteo Negri[1]
FBK-Irst[1], University of Trento[2]
Trento, Italy
[kouylekov,mehdad,negri]@fbk.eu

FONDAZIONE BRUNO KESSLER

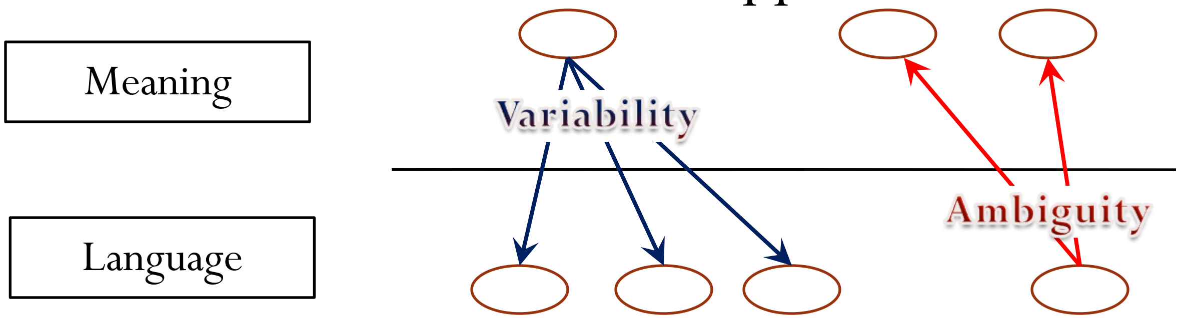
UNIVERSITÀ DEGLI STUDI DI TRENTO
Dipartimento di Ingegneria e Scienza dell'Informazione

Outline

- Recognizing Textual Entailment
- Lexical Knowledge in RTE
- Lexical Resources
  - WordNet
  - VerbOcean
  - Lin's dependency thesaurus
  - Lin's proximity thesaurus
- Mining Wikipedia
- Experiments
- Results
- Conclusion

Textual Entailment (TE) (Ido Dagan and Oren Glickman, 2004)

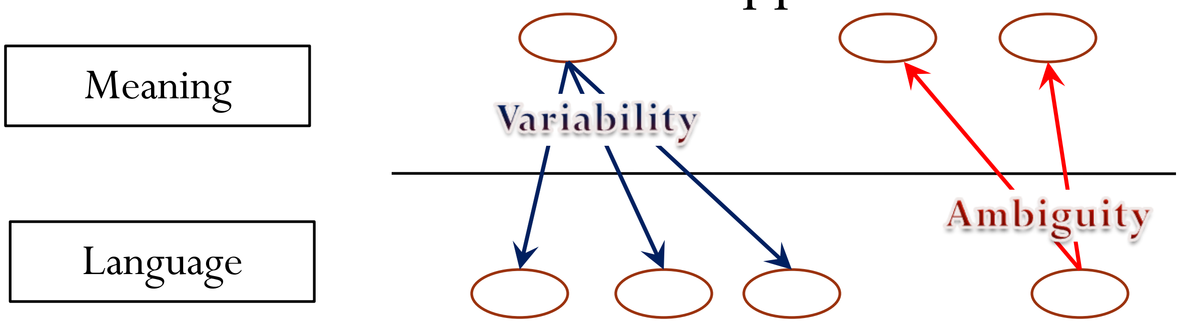
- Text applications require *semantic* inference.
- TE as a common framework for applied semantics.



- Definition: a text T entails a hypothesis H if, typically, a human reading T would infer that H is most likely true.

Textual Entailment (TE) (Ido Dagan and Oren Glickman, 2004)

- Text applications require *semantic* inference.
- TE as a common framework for applied semantics.



- Definition: a text T entails a hypothesis H if, typically, a human reading T would infer that H is most likely true.

YES

T: Profits doubled to about $1.8 billion.
H: Profits grew to nearly $1.8 billion.

NO

T: Time Warner is the world's largest media and Internet company.
H: Time Warner is the world's largest company.

Lexical Knowledge in RTE - Importance

- Substantial agreement on the usefulness of some prominent resources, including:
  - WordNet (Fellbaum, 1998)
  - eXtendedWordNet (Moldovan and Novischi, 2002)
  - Dependency and proximity thesauri (Lin, 1998)
  - VerbOcean (Chklovski and Pantel, 2004).
  - Wikipedia
  - FrameNEt
- Mirkin et al. (Mirkin et al., 2009):
  I. Most widely used resources for lexical knowledge (e.g. WordNet) allow for limited recall figures.
  II. Resources built considering distributional evidence (e.g. Lin's Dependency and Proximity thesauri) are suitable to capture more entailment relationships.
  III. The application of rules in inappropriate contexts severely impacts on performance.

Motivating Examples

pass away → die

T: Everest summiter David Hiddleston has **passed away** in an avalanche of Mt. Tasman.

H: A person **died** in an avalanche.

Begin → start

T: El Nino usually **begins** in December and lasts a few months.

H: El Nino usually **starts** in December.

European Union → EU

T: There are currently eleven (11) official languages of the **European Union** in number.

H: There are 11 official **EU** languages.

Lexical Entailment Rules
(Kouylekov and Magnini, 2006)

- Creation of repositories of lexical entailment rules.
- Each rule has a left hand side (W_T) and a right hand side (W_H).
- Associated to a probability: Pr ($W_T \rightarrow W_H$)
  - Eg. : [phobia ➔ disorder]

    - T: Agoraphobia means fear of open spaces and is one of the most common **phobias**.
    - H: Agoraphobia is a widespread **disorder**.

Rule Extraction - I

- **WordNet** rules: given a word w_1 in T, a new rule [$w_1 \rightarrow w_2$] is created for each word w_2 in H that is a synonym or an hypernym of w_1.
- **VerbOcean** rules: given a verb v_1 in T, a new rule [$v_1 \rightarrow v_2$] is created for each verb v_2 in H that is connected to v_1 by the [stronger-than] relation (i.e. when [v_1 stronger-than v_2]).
- **Lin Dependency/Proximity** Similarity rules are collected from the dependency and proximity based similarities described in (Lin, 1998).
  - Empirically estimate a relatedness threshold over training data to filter out all the pairs of terms featuring low similarity.

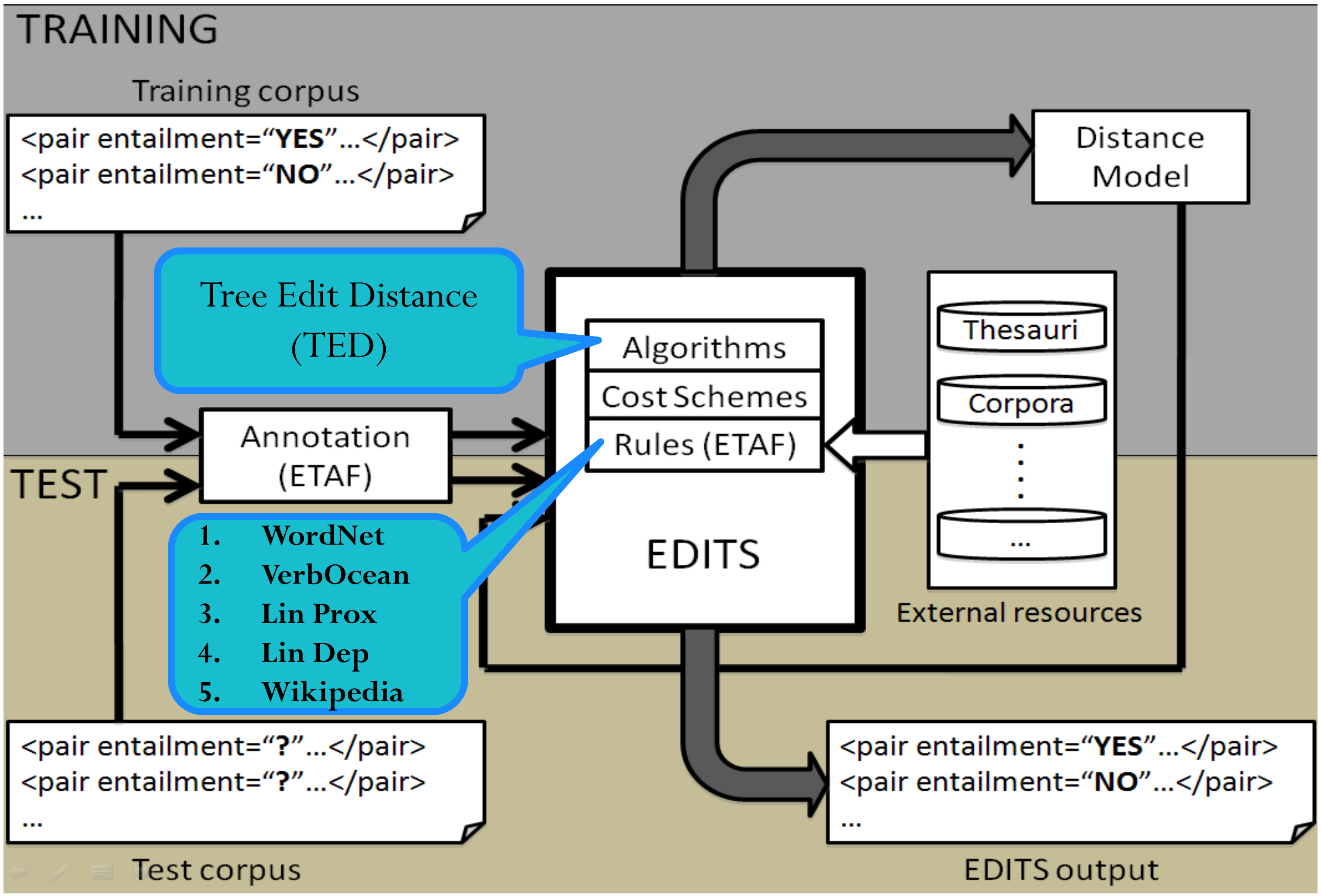
Rule Extraction – Mining Wikipedia

- Advantage:
  - Coverage: more than 3.000.000 articles with updated NE.
  - Context sensitivity: allows to consider the context in which rule elements tend to appear.
- Approach: Latent Semantic Analysis (LSA) score over Wikipedia between all possible word pairs that appear in the T-H pairs of an RTE dataset.
  - jLSI (java Latent Semantic Indexing)[1]
  - 200,000 most visited Wikipedia articles.
  - Empirically estimate a relatedness threshold over training data to filter out all the pairs of terms featuring low similarity.

1- http://tcc.itc.it/research/textec/tools-resources/jLSI.html

Experiments - I

- EDITS (Edit Distance Textual Entailment Suite)[2]

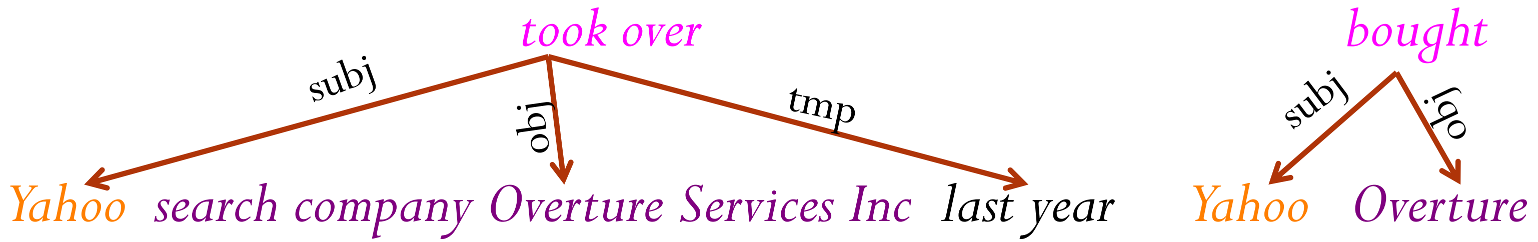


2- Kouylekov, Negri : An Open-source Package for Recognizing Textual Entailment. ACL 2010 Demo

TED for RTE

T: *Yahoo took over search company Overture Services Inc last year*

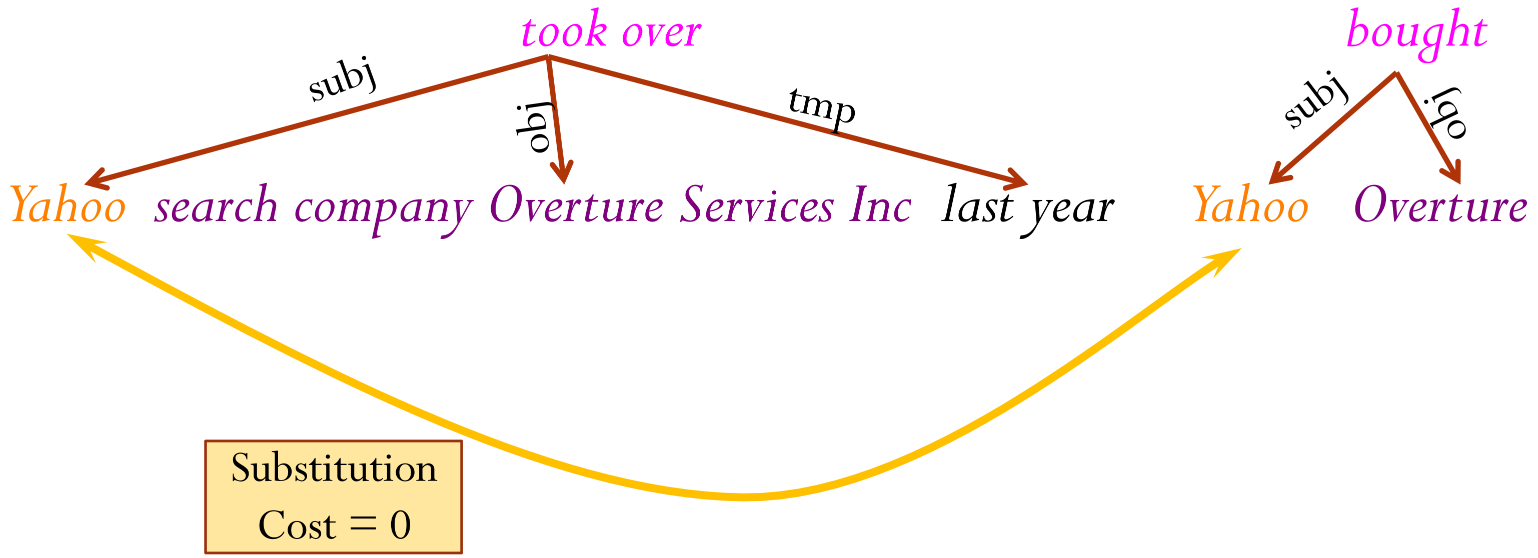
H: *Yahoo bought Overture*



TED for RTE

T: *Yahoo took over search company Overture Services Inc last year*

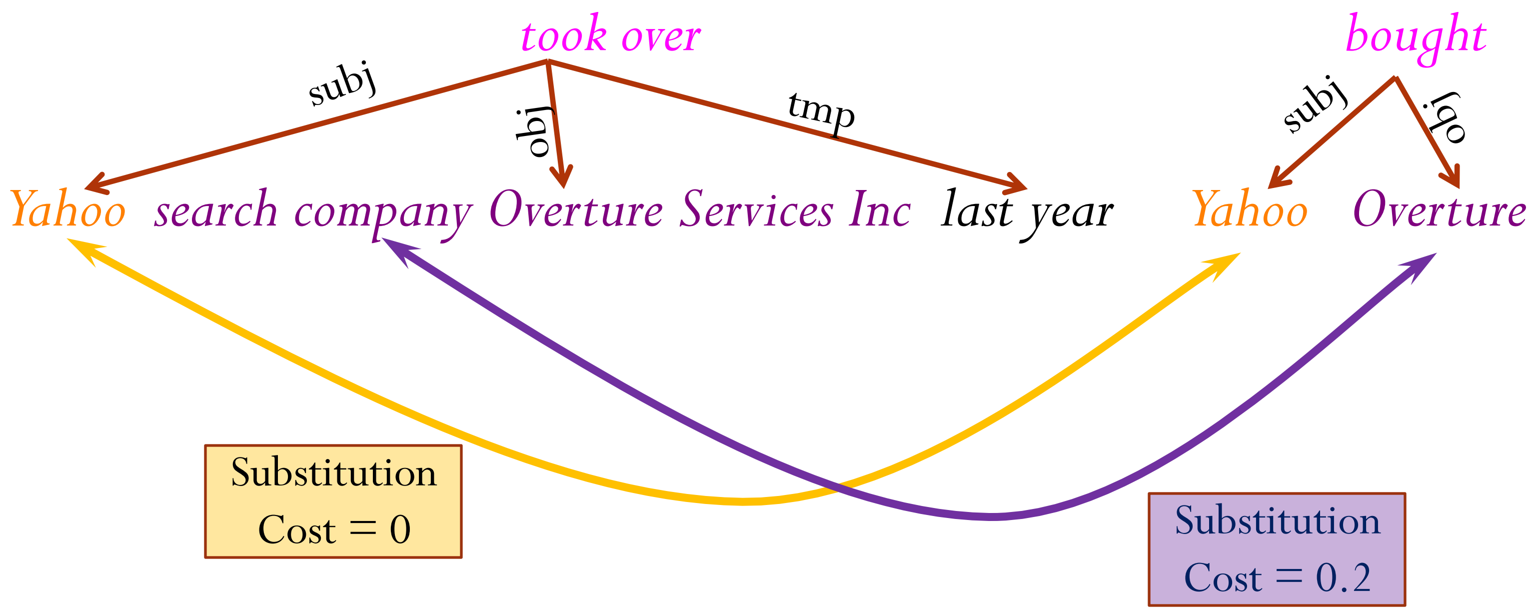
H: *Yahoo bought Overture*



TED for RTE

T: *Yahoo took over search company Overture Services Inc last year*

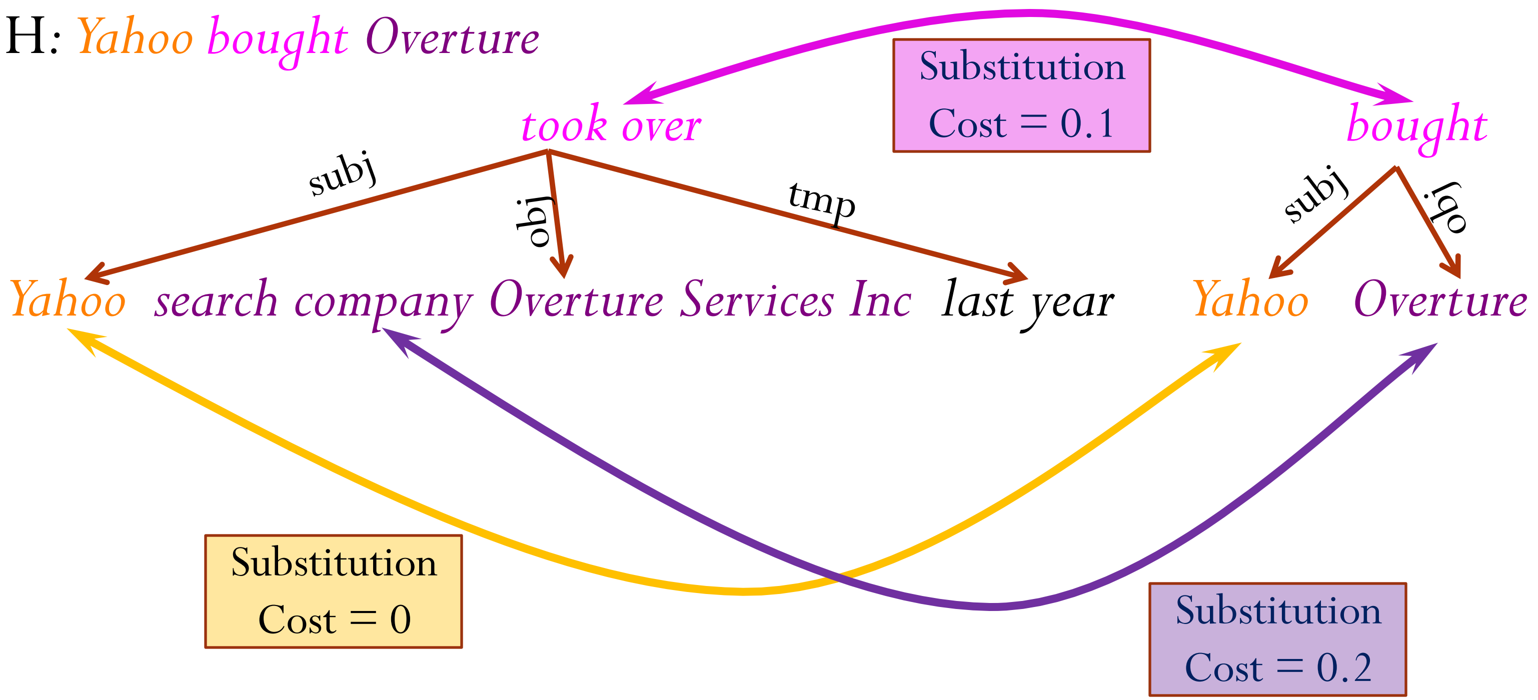
H: *Yahoo bought Overture*



TED for RTE

T: *Yahoo took over search company Overture Services Inc last year*

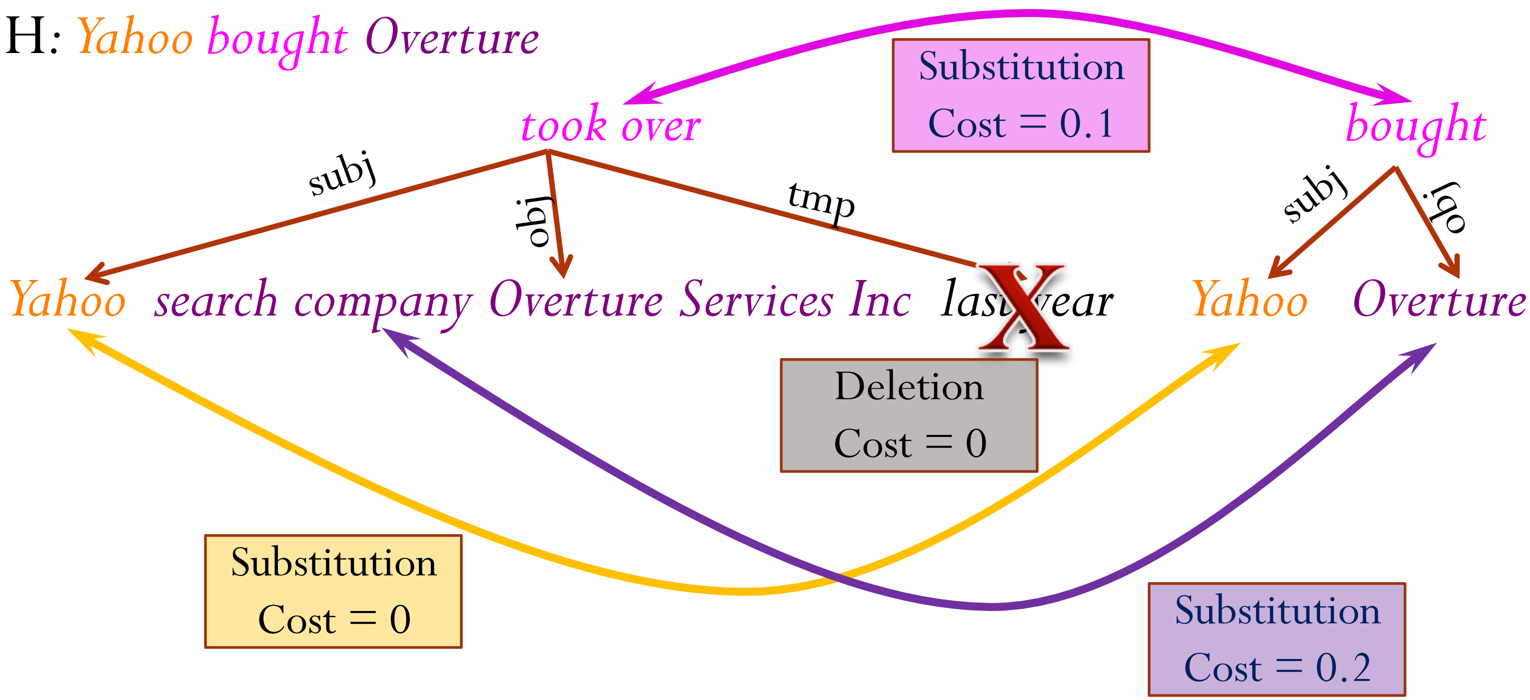
H: *Yahoo bought Overture*



TED for RTE

T: *Yahoo took over search company Overture Services Inc last year*

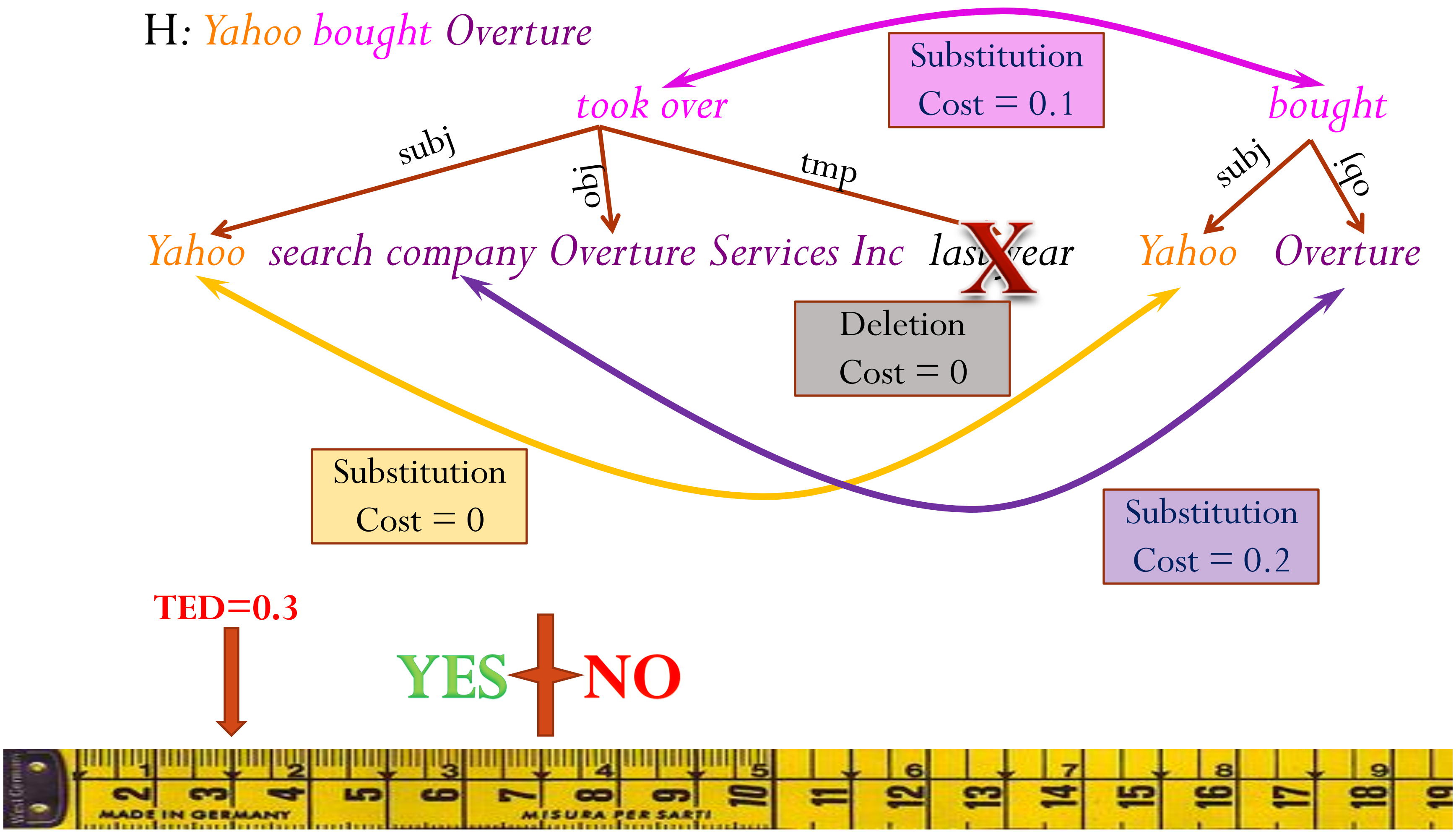
H: *Yahoo bought Overture*



TED for RTE

T: *Yahoo took over search company Overture Services Inc last year*

H: *Yahoo bought Overture*



Experiments

- Dataset: RTE5 (the most recent RTE data)
- Rule repositories

1. WIKI: Original 199217 rules extracted, 58278 retained
2. WN: 1106 rules
3. VO: 192 rules
4. DEP: 5432 rules extracted from Lin's dependency thesaurus, 2468 rules retained
5. PROX: 8029 rules extracted from Lin's proximity thesaurus, 236 retained

Results

Baseline (No rules): Dev: 58.3 Test: 56

RTE5	VO		WN		PROX		DEP		WIKI	
	DEV	TEST	DEV	TEST	DEV	TEST	DEV	TEST	DEV	TEST
Acc.	**61.8**	**58.8**	**61.8**	**58.6**	**61.8**	**58.8**	**62**	**57.3**	**62.6**	**60.3**

+ 0.5-1% +1.5-2%

- ✓ Performance improvement
- ✓ Example of Wiki rules:
  - ➢ Apple → Macintosh
  - ➢ Iranian → IRIB

Coverage Analysis

- ✓ Increasing the coverage using a context sensitive approach in rule extraction, may result in a better performance in the RTE task.
- ✓ Count the number of pairs in the RTE-5 data which contain rules present in the WordNet, VerbOcean, Lin Dependency/Proximity, and Wikipedia repositories.

Rules	VO		WN		PROX		DEP		WIKI	
	Extracted	Retained	Extracted	Retained	Extracted	Retained	Extracted	Retained	Extracted	Retained
Coverage %	**0.08**	**0.08**	**0.4**	**0.4**	**3**	**0.09**	**2**	**1**	**83**	**24**

Conclusion

- Experiments with lexical entailment rules from Wikipedia.
- Aim to maximizing two key features:
  - ✓ Coverage: the proportion of rules successfully applied
  - ✓ Context sensitivity: the proportion of rules applied in appropriate contexts
- Improvement on RTE5 dataset using Wikipedia rules.
- Very high coverage in comparison with other resources.
- Noise (low accuracy) is not always harmful.
- Flexible approach for extracting entailment rules regardless of language dependency.

Challenges and Remarks

- Performance increase is lower than expected.
  - The difficulty of exploiting lexical information in TED algorithm.
  - Valid and reliable rules that could be potentially applied to reduce the distance between T and H are often ignored because of the syntactic constraints imposed.
  - Some rules were applied to the negative examples.
- Future work:
  - Definition of more flexible algorithms.
    - Capable of exploiting the full potential offered by Wikipedia rules.
  - Development of other methods for extracting entailment rules from Wikipedia.

LSA (more on computation)

- SVD (Singular Value Decomposition)

$$A_{m \times n} = U_{m \times r} \Sigma_{r \times r} V^{T}_{r \times n}$$

where

$$\Sigma = diag(\sigma_1, \ldots, \sigma_r)$$

$$\sigma_1 \geq \sigma_1 \geq \ldots \geq \sigma_r \geq 0 \text{ and } r = \min\{m, n\}$$

A: weighted matrix of term frequencies in a collection of text

U: matrix of term vectors

∑: diagonal matrix containing the singular value of A

V: matrix of document vectors