Paraphrase Recognition Using Machine Learning to Combine Similarity Measures

Prodromos Malakasiotis

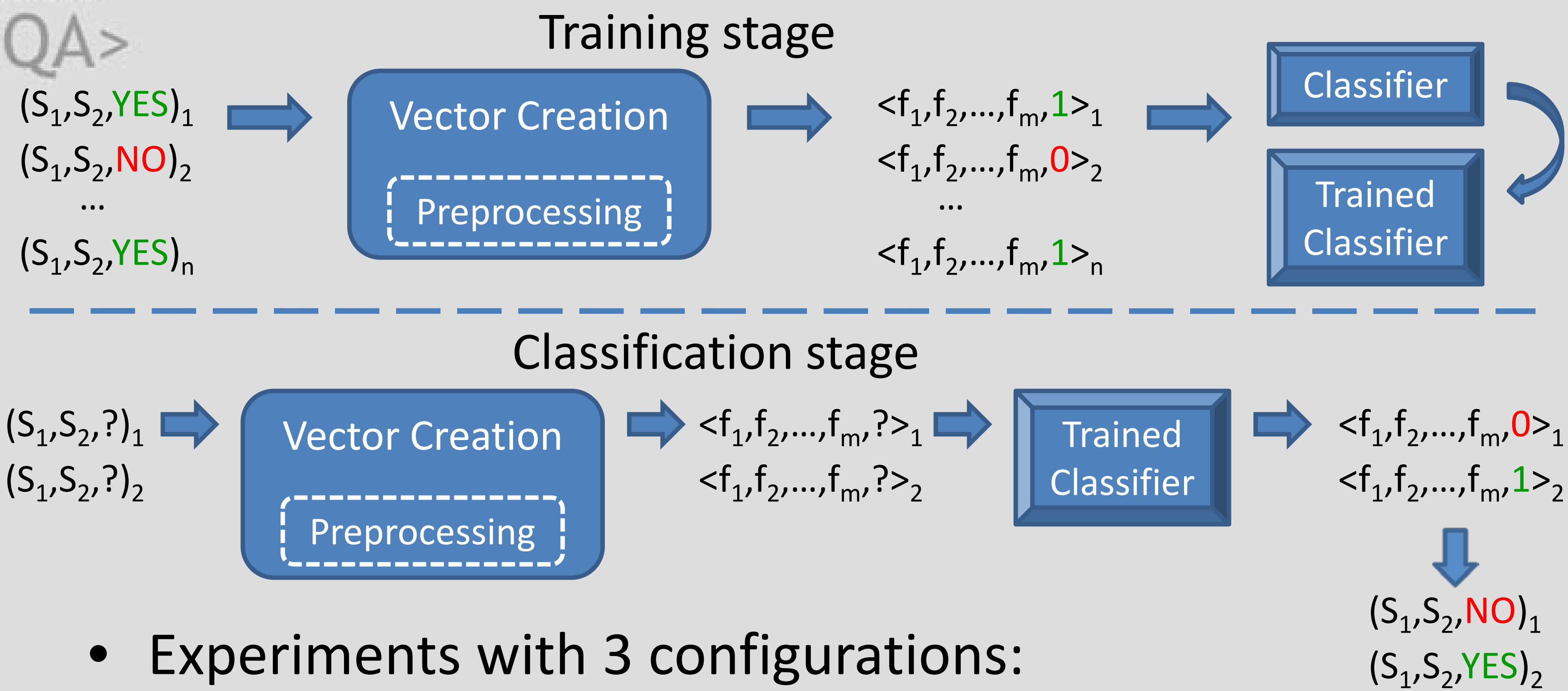
Department of Informatics

Athens University of Economics and Business

Paraphrase Recognition

- Given a pair of phrases, sentences, or patterns [S_1 , S_2] decide if they are paraphrases, i.e., if they have (almost) the same meaning.
  - “X is the writer of Y” ≈ “X wrote Y” ≈ “X is the author of Y”
- Related to, but not the same as textual entailment.
  - “Athens is the capital of Greece” $\models$ “Athens is located in Greece”, but not the reverse.
- Paraphrasing can be seen as bidirectional textual entailment.

Paraphrase recognition with Machine Learning



- Experiments with 3 configurations:
  - INIT, INIT+WN,INIT+WN+DEP

INIT Configuration

- The input pairs [S_1 , S_2] are represented as vectors of similarity scores measured on 4 forms of [S_1 , S_2]:
  - (1) words, (2) stems, (3) POS-tags, (4) soundex codes
- 9 similarity measures, applied to the 4 forms:
  - Levenshtein (edit distance), Jaro-Winkler, Manhattan, Euclidean distance, cosine similarity, n-gram, matching coefficient, Dice, and Jaccard coefficient (see paper).
  - Similarities are measured in terms of tokens.

Partial Matching Features

S_1 While Bolton apparently fell and was immobilized, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

S_2 After the other inmate fell, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

Avg: 0.64

S_1 $\boxed{W_1 W_2 W_3 W_4 W_5 W_6 W_7 W_8} W_9 W_{10}$

S_2 $\boxed{W'_1 W'_2 W'_3 W'_4 W'_5 W'_6 W'_7 W'_8}$

- Find S_1's (longer sentence) part that is most similar to S_2 (shorter sentence) using a sliding window:
  - At each step, calculate the average of 9 similarity scores.
- Use the highest average (Avg) and the 9 scores it was computed from as additional features in INIT.
- Do this for words, stems, POS-tags, and soundex codes.

Partial Matching Features

S_1 While Bolton apparently fell and was immobilized, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

S_2 After the other inmate fell, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

Avg: 0.71

S_1 $W_1 \boxed{W_2 W_3 W_4 W_5 W_6 W_7 W_8 W_9} W_{10}$

S_2 $\boxed{W'_1 W'_2 W'_3 W'_4 W'_5 W'_6 W'_7 W'_8}$

- Find S_1's (longer sentence) part that is most similar to S_2 (shorter sentence) using a sliding window:
  - At each step, calculate the average of 9 similarity scores.
- Use the highest average (Avg) and the 9 scores it was computed from as additional features in INIT.
- Do this for words, stems, POS-tags, and soundex codes.

Partial Matching Features

S_1 While Bolton apparently fell and was immobilized, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

S_2 After the other inmate fell, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

Avg: 0.82

$$S_1 \quad W_1 W_2 \boxed{W_3 W_4 W_5 W_6 W_7 W_8 W_9 W_{10}}$$

$$S_2 \quad \boxed{W'_1 W'_2 W'_3 W'_4 W'_5 W'_6 W'_7 W'_8}$$

- Find S_1's (longer sentence) part that is most similar to S_2 (shorter sentence) using a sliding window:
  - At each step, calculate the average of 9 similarity scores.
- Use the highest average (Avg) and the 9 scores it was computed from as additional features in INIT.
- Do this for words, stems, POS-tags, and soundex codes.

A Partial Matching Example

Initial Avg: 0.76

While Bolton apparently fell and was immobilized, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

Avg: 0.68

After the other inmate fell, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

While Bolton apparently fell and was immobilized, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

Avg: 0.72

After the other inmate fell, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

While Bolton apparently fell and was immobilized, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

Avg: 0.82

After the other inmate fell, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

INIT+WN Configuration

- The same as INIT, but:
  - It treats words from S_1 and S_2 that are synonyms (in WordNet) as identical.

Fewer than a dozen FBI agents were dispatched to secure and analyze evidence.

Fewer than a dozen FBI agents will be sent to Iraq to secure and analyze evidence of the bombing.

INIT's Avg: 0.73

INIT+WN's Avg: **0.78**

INIT+WN+DEP Configuration

- Same as INIT+WN, but:
  - 3 additional features that measure dependency grammar similarity between S_1 and S_2:

$$R_1 = \frac{|\,common\ dependencies\,|}{|\,S_1\ dependencies\,|}$$

$$R_2 = \frac{|\,common\ dependencies\,|}{|\,S_2\ dependencies\,|}$$

$$F_{R_1,R_2} = \frac{2 \cdot R_1 \cdot R_2}{R_1 + R_2}$$

INIT+WN+DEP: negative example

The dollar was at 116.92 yen against the yen, flat on the session, and at 1.2891 against the Swiss franc, also flat.

det(dollar-2, The-1)
nsubj(flat-25, dollar-2)
…
dep(at-4, against-7)
det(yen-9, the-8)
…
det(session-14, the-13)
pobj(on-12, session-14)
…

The dollar was at 116.78 yen JPY, virtually flat on the session, and at 1.2871 against the Swiss franc CHF, down 0.1 percent.

det(dollar-2, The-1)
nsubj(was-3, dollar-2)
…
det(session-14, the-13)
prep_on(flat-11, session-14)
…
dep(at-17, against-19)
det(CHF-23, the-20)
…

Avg = 0.72
$R_1 = 0.14$
$R_2 = 0.16$
$F_{R_1,R_2} = 0.15$

INIT+WN+DEP: positive example

Last week the power station's US owners, AES Corp, walked away from the plant after banks and bondholders refused to accept its financial restructuring offer.

amod(week-2, Last-1)
tmod(walked-13, week-2)
…
prt(walked-13, away-14)
…

The news comes after Drax's American owner, AES Corp. AES.N, last week walked away from the plant after banks and bondholders refused to accept its restructuring offer.

det(news-2, The-1)
…
amod(week-18, last-17)
dep(walked-19, week-18)
…
prt(walked-19, away-20)
…

Avg= 0.71
R_1 = 0.52
R_2 = 0.59
F_{R_1,R_2} = 0.55

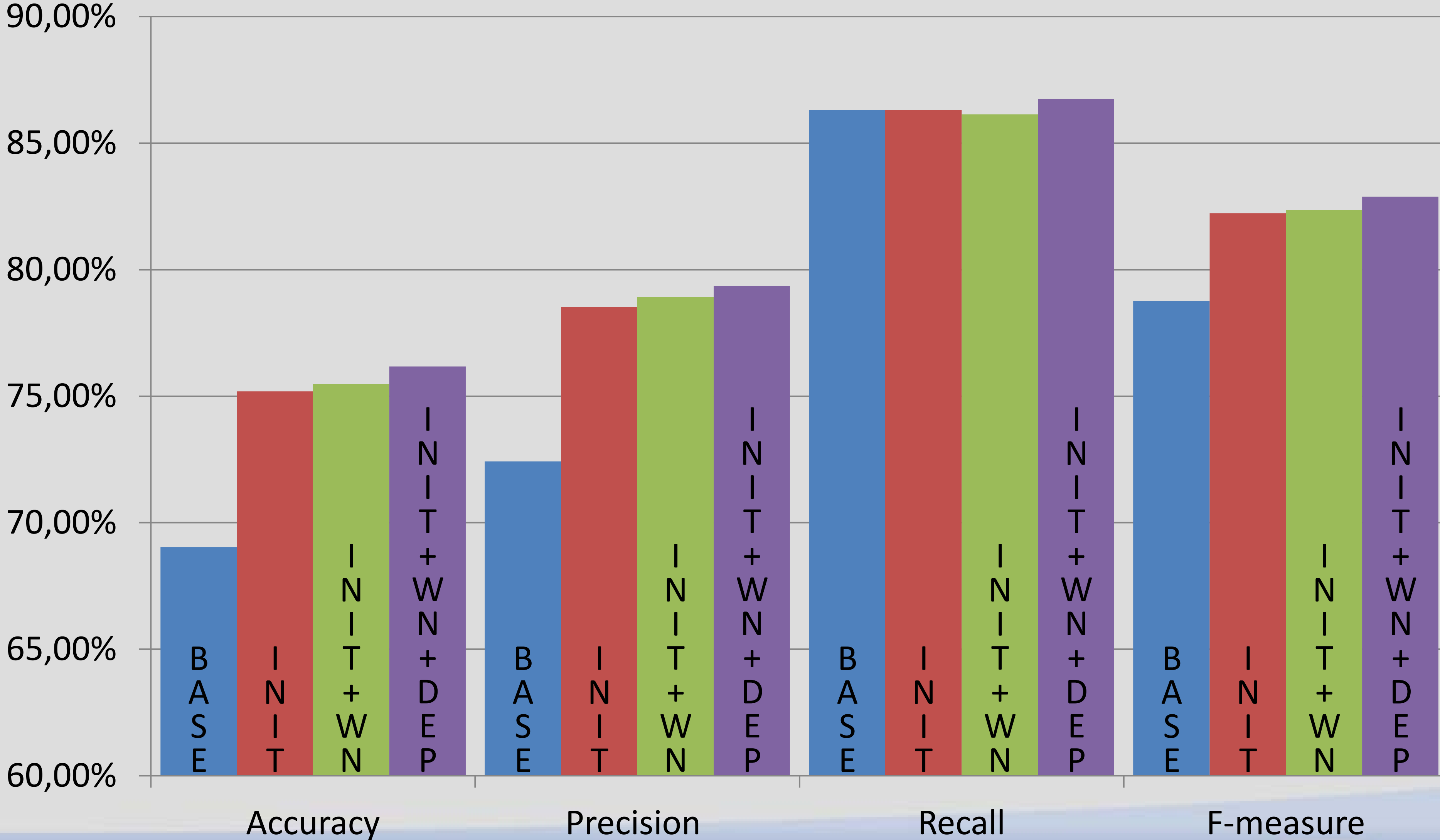
Feature Selection

- Start with an empty feature set.
- Gradually add features:
  - Form new feature sets by adding one feature.
  - Measure the predictive power of the new sets.
  - Keep the best new feature set(s).
  - Tried both hill-climbing and beam-search.
- A lot of redundancy in the full feature set.
  - Feature selection leads to competitive results with much fewer features (10 instead of 136).
- But the full feature set leads to better results.

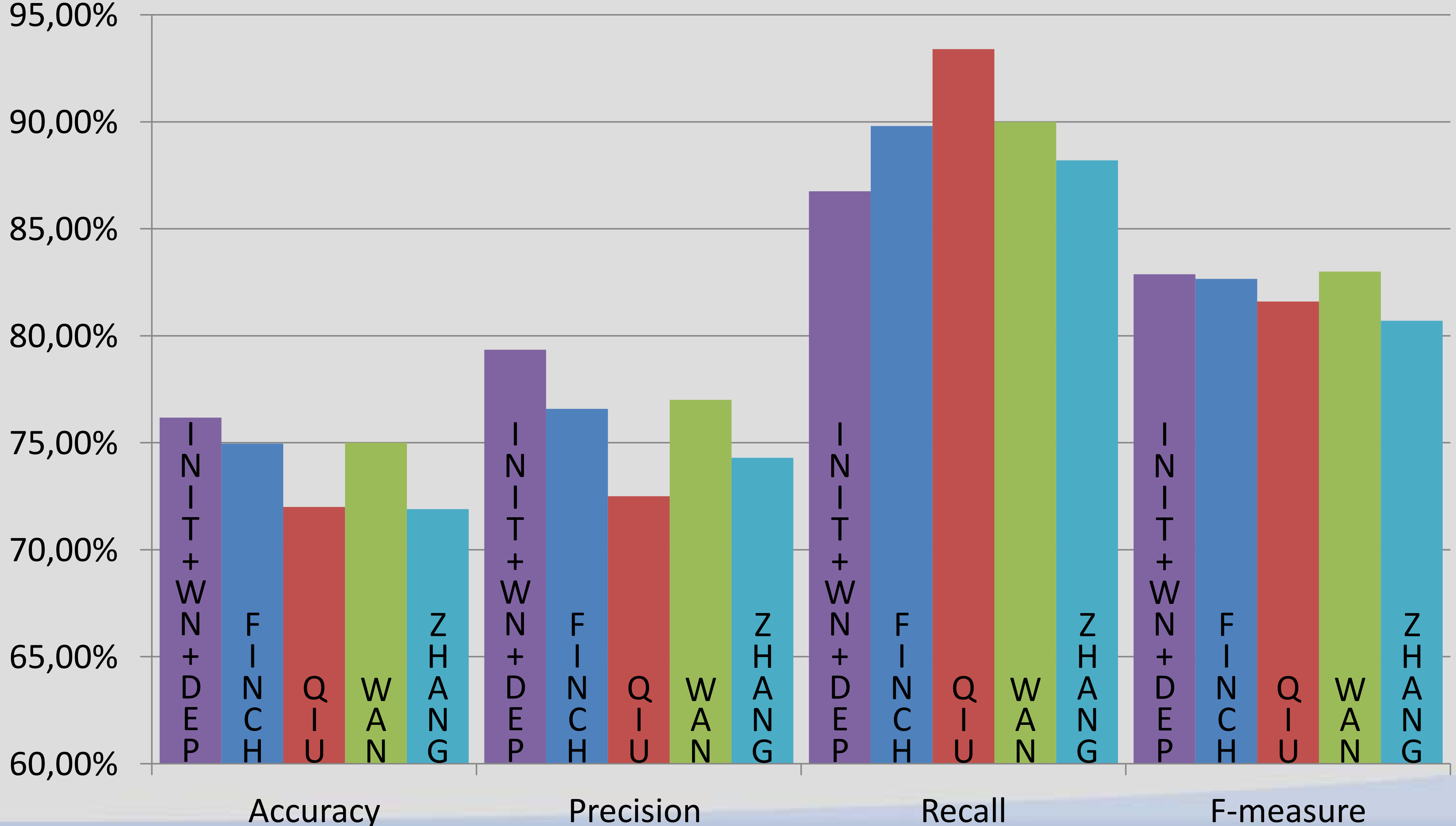
Experiments

- Microsof Research (MSR) Paraphrase Corpus:
  - 5,801 pairs of sentences evaluated by judges.
  - 4,076 training pairs.
  - 1,725 testing pairs.
- Baseline (BASE):
  - Use a threshold on edit distance to decide if a pair is positive (paraphrases) or negative.
  - The threshold is tuned on the training pairs.

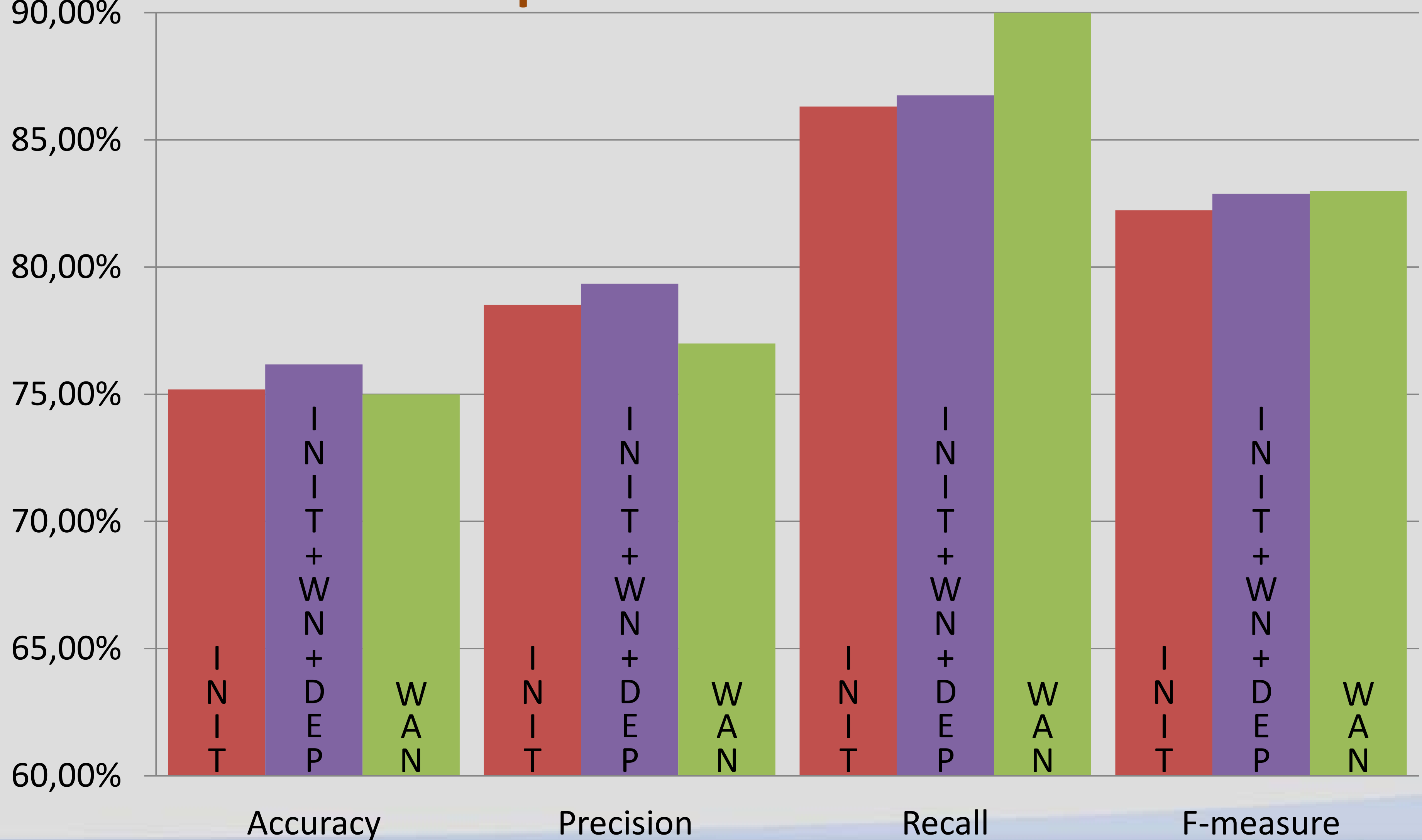
Results on MSR corpus



Best known results



INIT performs well too!



Conclusions

- INIT is competitive to the best known systems, using fewer resources.
  - Useful for languages where WordNet, reliable dependency parsers etc. are unavailable.
- INIT+WN and INIT+WN+DEP perform even better, but they require more resources and the improvement is small.
  - The differences may be small because of the high lexical overlap of the paraphrases in the MSR corpus.

Thank you!

Questions?