Anaphora resolution and coreference: three perennial questions

Ruslan Mitkov

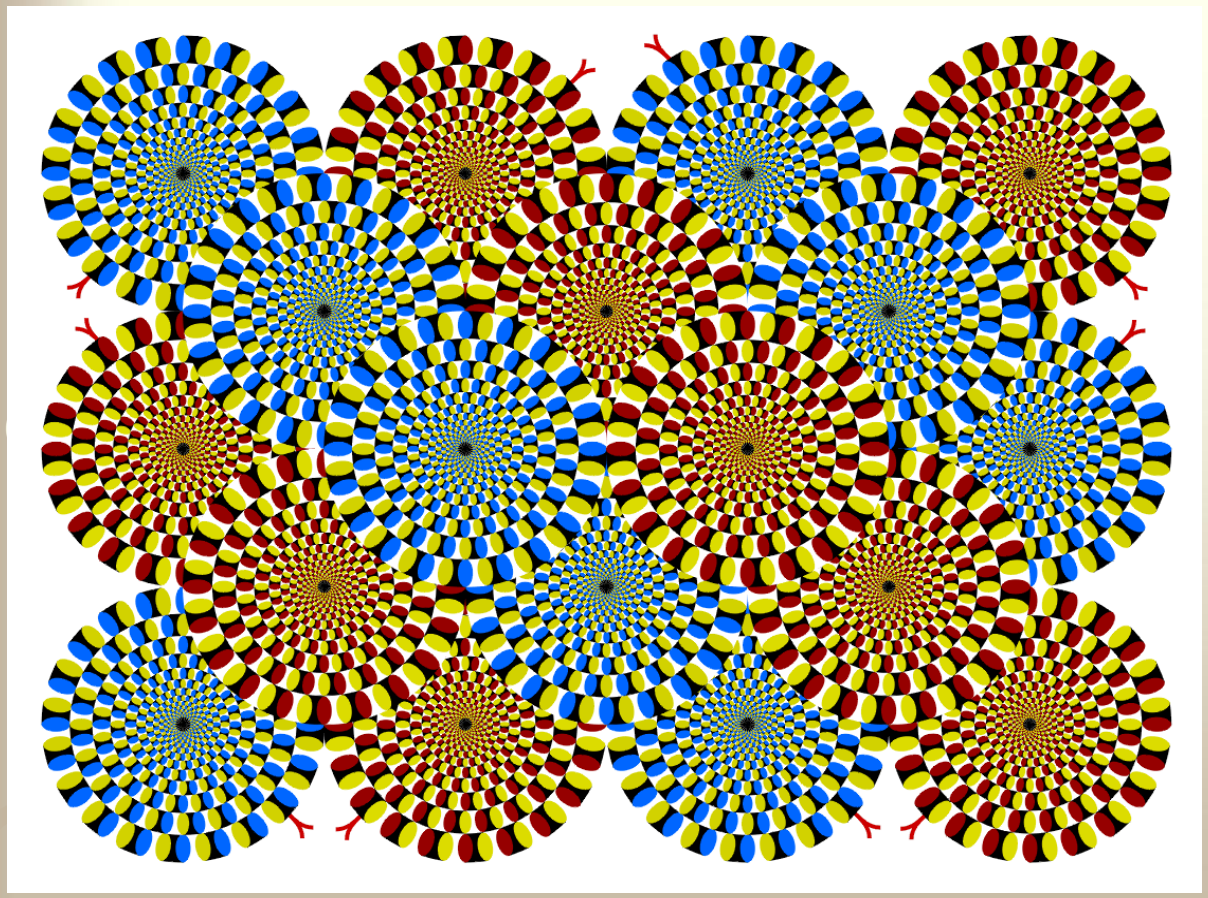
Research Group in Computational Linguistics

University of Wolverhampton

Anaphora resolution and coreference: three perennial questions

1. Are (automatic) anaphora resolution and coreference resolution beneficial to NLP applications?
2. Do we know how to evaluate anaphora resolution algorithms?
3. Which are the coreferential links most difficult to resolve?

Outline of the presentation



- Terminological notes
- The impact of anaphora and coreference resolution on NLP applications
- Evaluation of anaphora resolution
- Coreference links and cognitive efforts on readers

Anaphora vs. coreference

- Anaphora and coreference are not identical phenomena
- Anaphora which is not coreference: identity of sense anaphora
- The man who gave his paycheck to his wife was wiser than the man who gave it to his mistress
- Coreference which is not anaphora:
- Cross-document coreference

Anaphora (and coreference) resolution

- Anaphora resolution: tracking down the antecedent of an anaphor
- Coreference resolution: identification of all coreference classes (chains).

Anaphora and coreference: 3 perennial questions

1. Are (automatic) anaphora resolution and coreference resolution beneficial to NLP applications?
2. Do we know how to evaluate anaphora resolution algorithms?
3. Which are the coreferential links most difficult to resolve?

Objectives of Study 1

- To integrate a pronoun resolution system (MARS) within 3 NLP applications (text summarisation, term extraction, text categorisation)
- To evaluate these applications with and without a pronoun resolution module
- To establish of impact of pronoun resolution on these NLP applications

Objectives of Study 2

- To integrate a coreference resolution system (BART) within 3 NLP applications (text summarisation, text categorisation, recognising textual entailment)
- To evaluate these applications with and without the coreference resolution module
- To establish of impact of coreference resolution on these NLP applications

Study 1

- Mitkov's knowledge-poor pronoun resolution algorithm (MARS'02 and MARS'06)
- Newspaper articles published in *New Scientist* (55 texts from BNC)
- Short enough to be manually annotated
- Suitable for all extrinsic evaluation tasks performed
- Articles manually categorised into six classes – "Being Human", "Earth", "Fundamentals", "Health", "Living World", and "Opinion"
- Caution: MARS was not specially tuned to these genres!

Evaluation data (2)

- 1,200 3rd person pronouns; over 48,000 words
- Very short and very long texts filtered out
- Annotation: PALinkA (Orasan, 2003)
- Several layers of annotations:
  - Coreference
  - Important sentences
  - Terms
  - Topics

Extrinsic evaluation

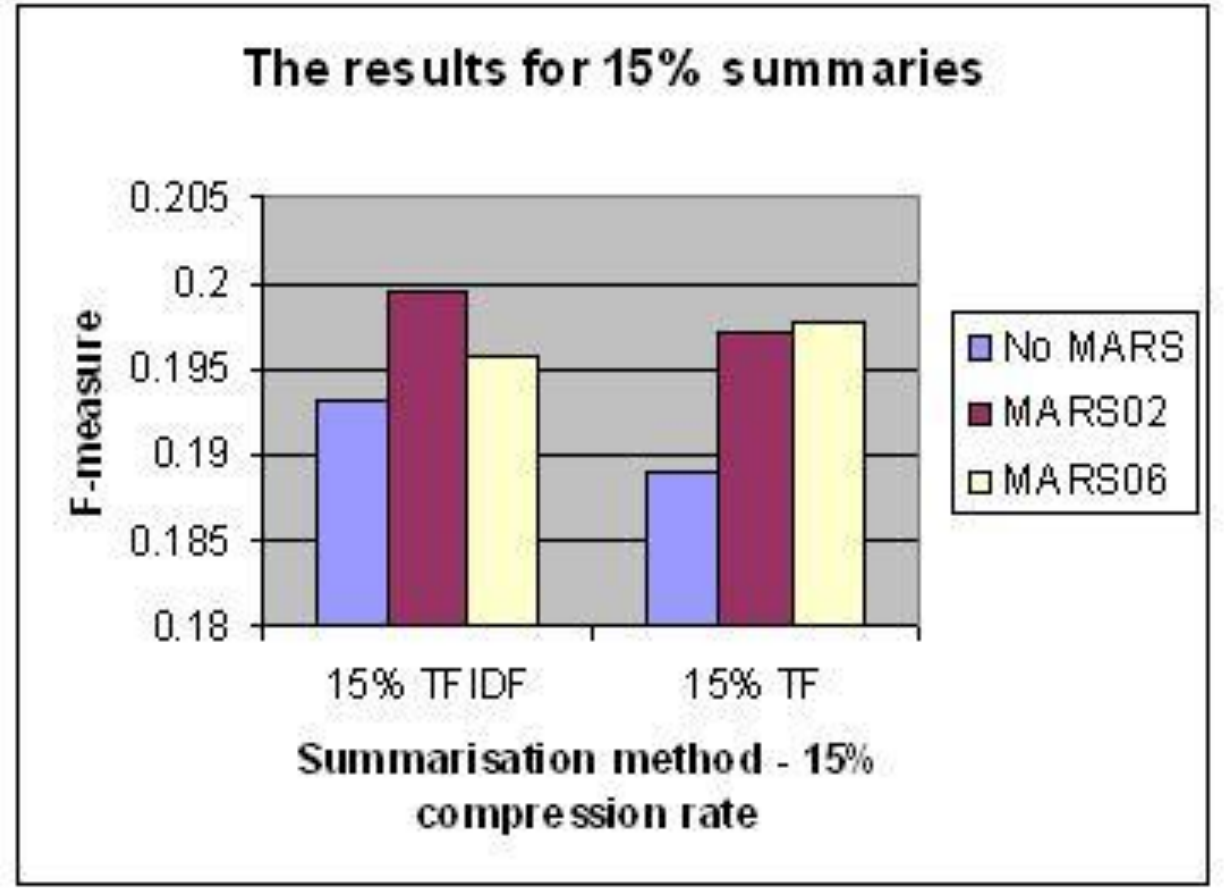
- Text summarisation
- Term extraction
- Text categorisation

Text summarisation

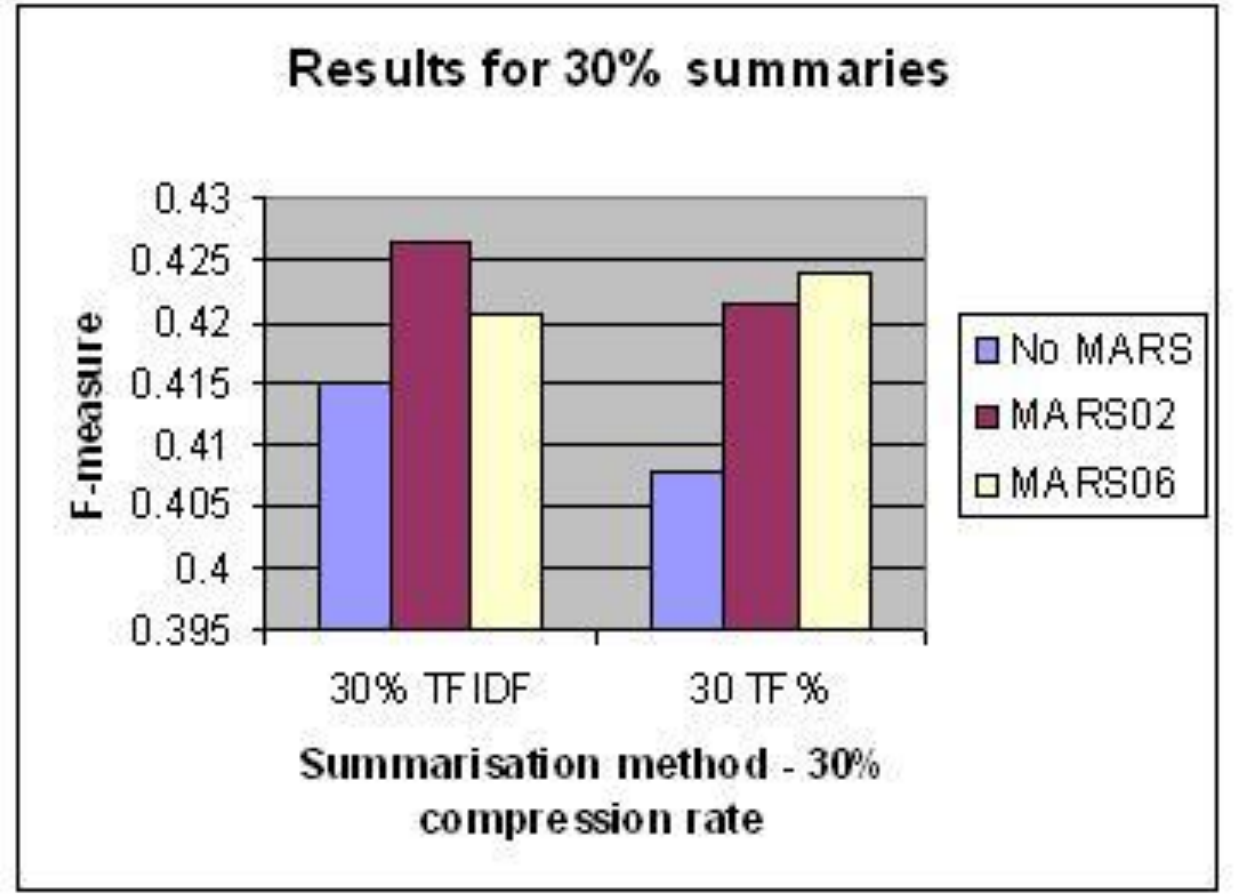
Summarisation

- Two term weighting methods investigated: term frequency and TF*IDF
- Evaluation measures: precision, recall and F-measure
- Evaluation performed for two (15% and 30%) compression rates

Summarisation (2)



Summarisation (3)

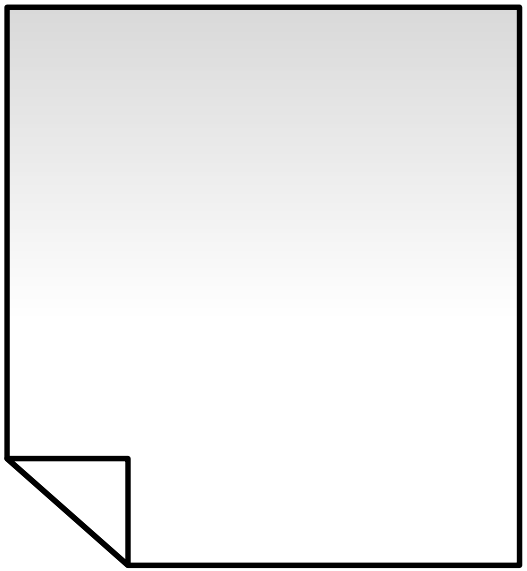


Summarisation (4)

- F-measure increases when anaphora resolution method employed
- Increase not statistically significant (T-test)
- Term frequency: results better for MARS'06
- TF.IDF: results better for MARS'02

Term extraction

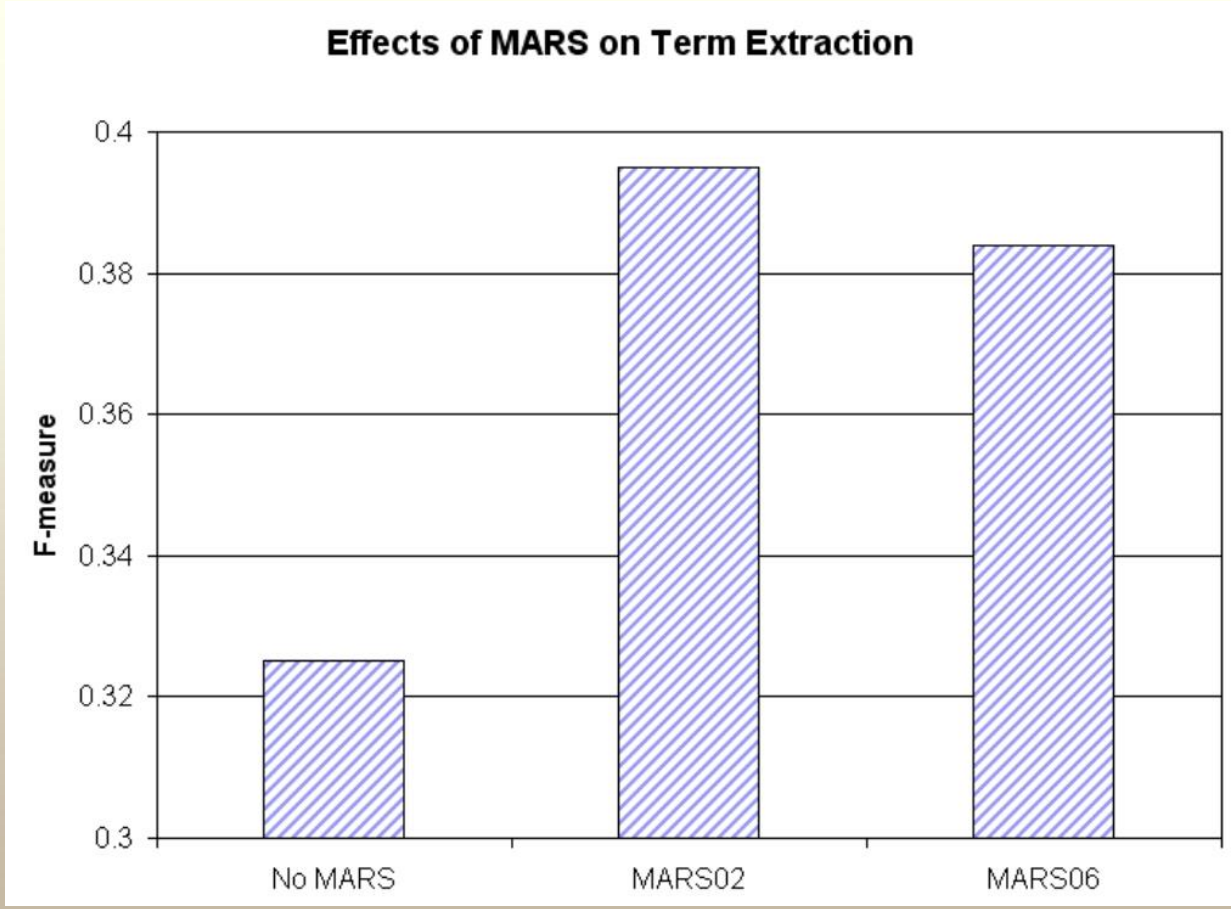
Natural language processing (NLP) is a field of computer science, artificial intelligence and linguistics concerned with the interactions between computers and human (natural) languages.



Term extraction

- Hybrid approach which combines statistical and lexical-syntactic filters in line with (Justeson and Katz 1986) and (Hulth 2003).
- Evaluation measures: precision, recall and F-measure.

Term extraction (2)



Term extraction (3)

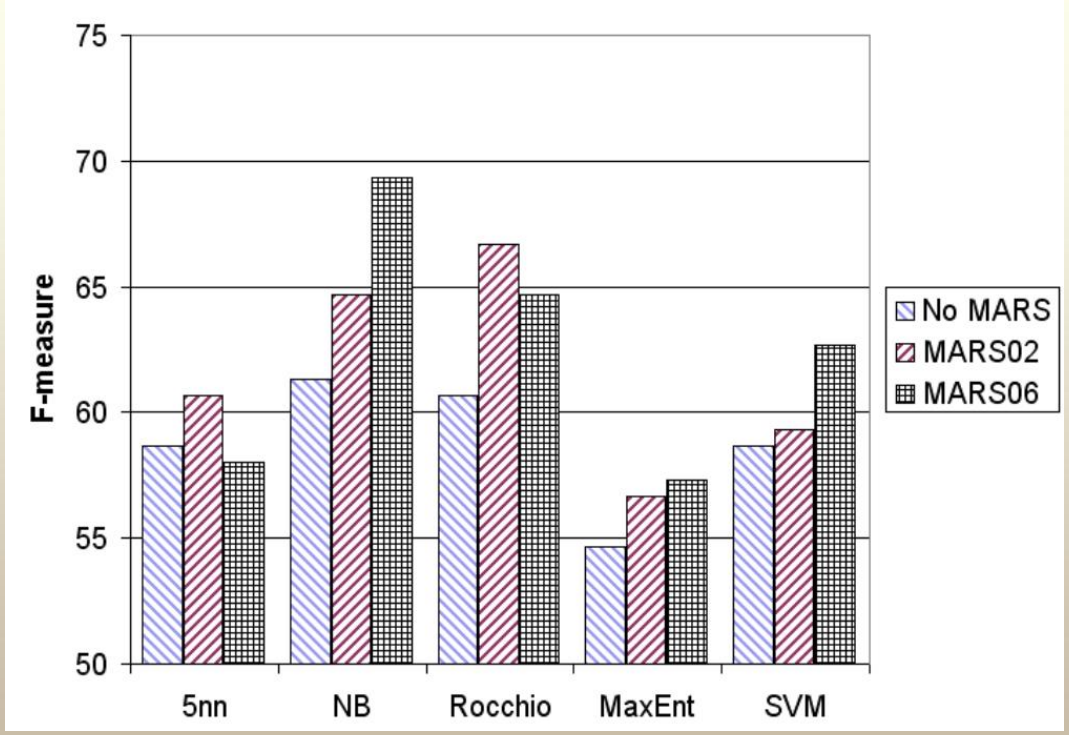
- F-measure increases when anaphora resolution method employed
- Increase not statistically significant (T-test)
- MARS'02 fares better in general
- MARS'02 improves both precision and recall
- MARS'06 improves mostly recall

Text categorisation

Text categorisation

- 5 different text classification methods: k nearest neighbours, Naïve Bayes, Rocchio, Maximum Entropy, and Support Vector Machines.
- Evaluation measures: precision, recall and F-measure

Text categorisation (2)



Text categorisation (3)

- F-measure increases in *most cases* when anaphora resolution method employed
- Increase not statistically significant for any of the methods

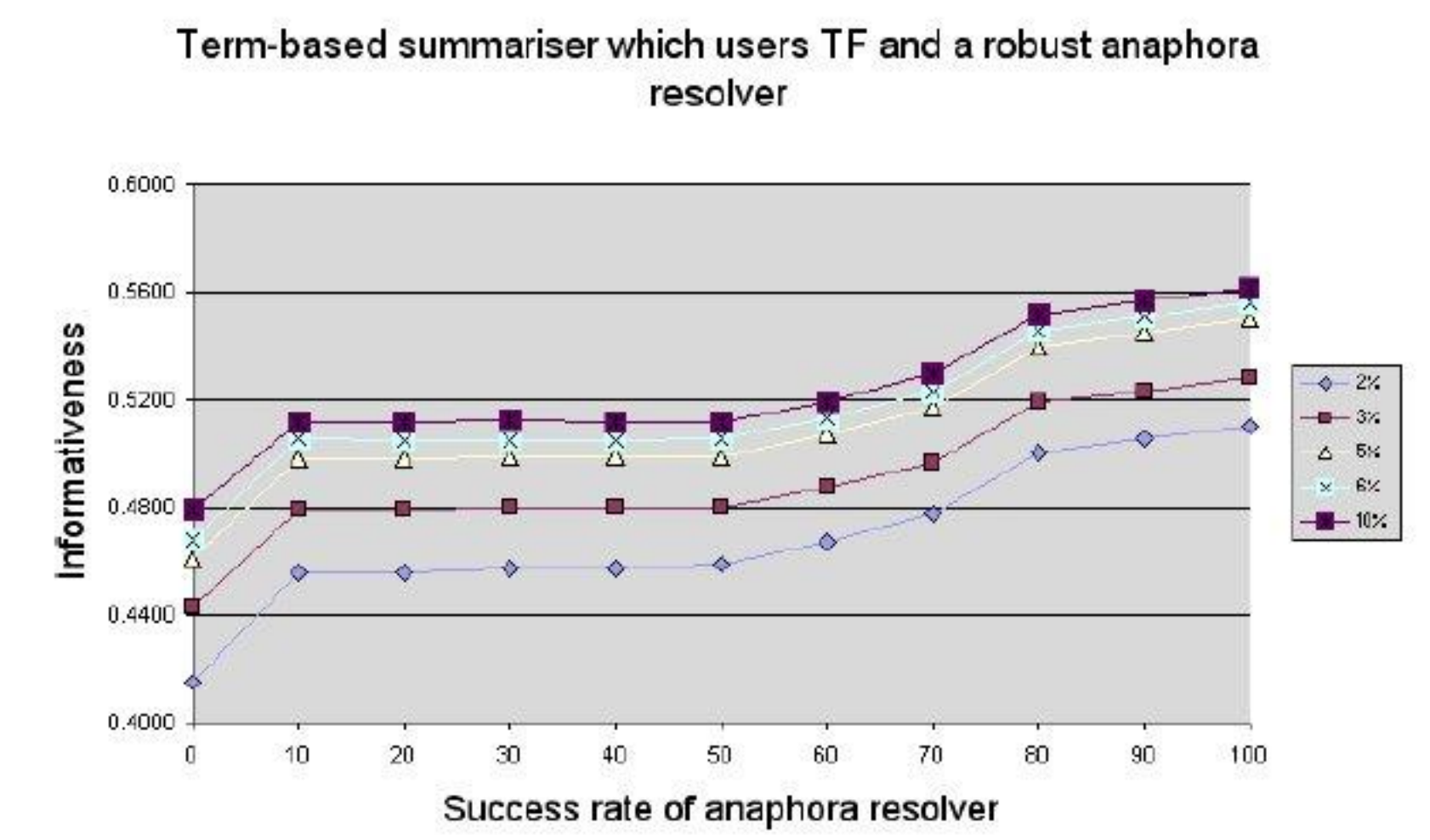
Discussion

- By and large deployment of MARS has positive but limited impact
- Would dramatic improvement in anaphora resolution lead to a marked improvement of NLP applications?

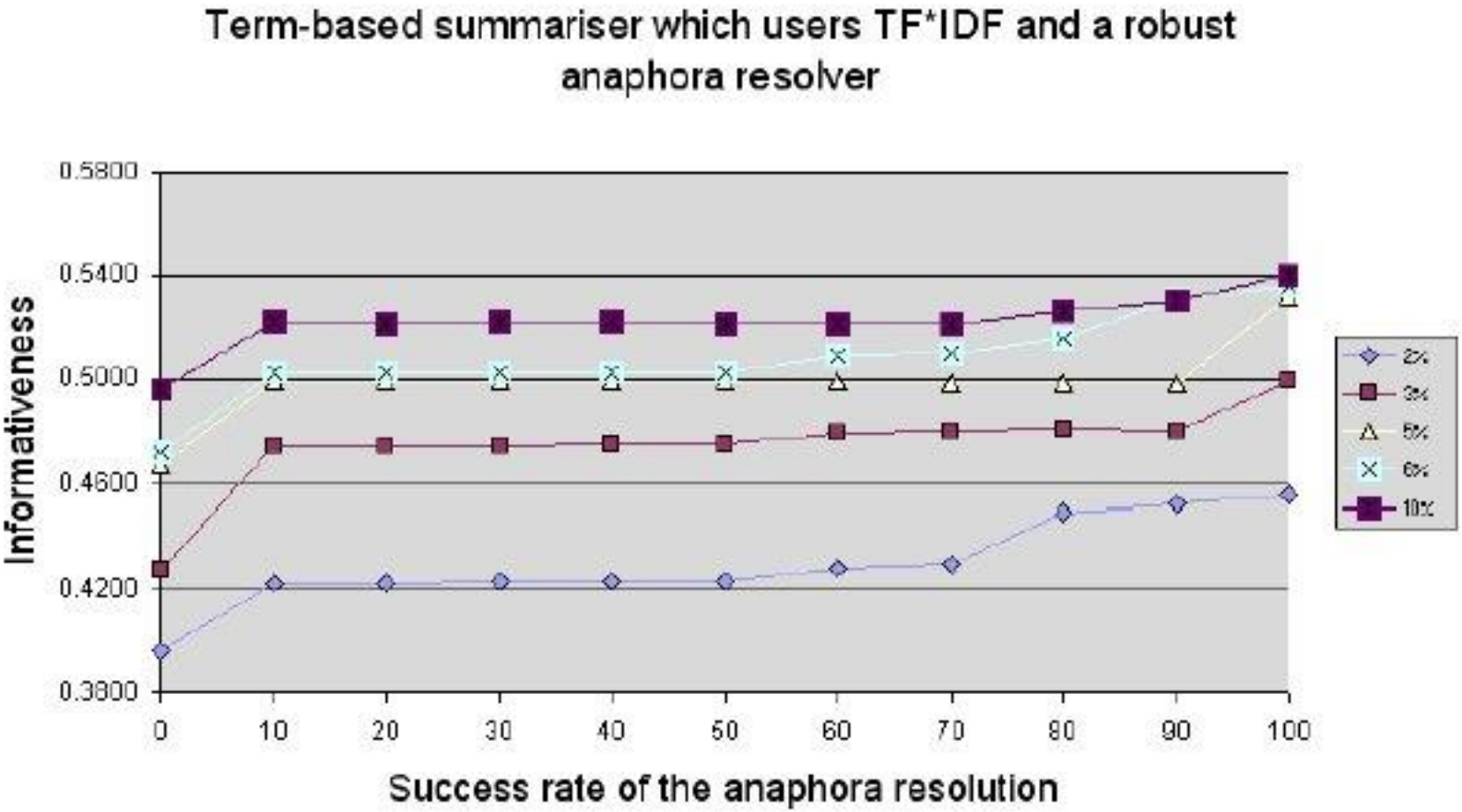
Would dramatic improvement in anaphora resolution lead to a marked improvement of NLP applications?

- Experiments on text summarisation (Orasan 2006)
- On a corpus of scientific articles anaphora resolution helps ….
  - TF summarisation if performance over 60-70%
  - TF.IDF summarisation if performance above 80%

Would dramatic improvement in anaphora resolution lead to a marked improvement of NLP applications? (2)



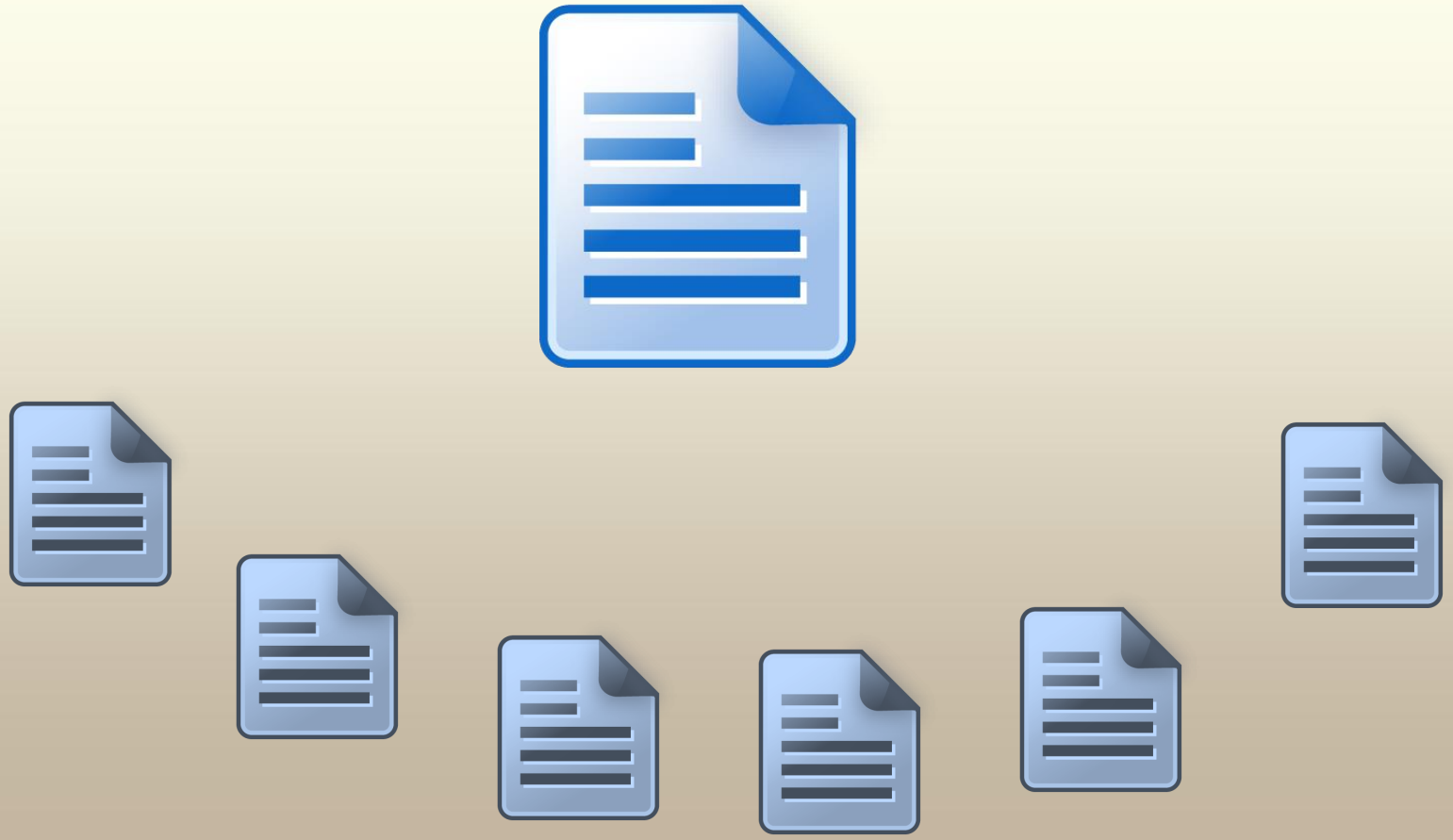
Would dramatic improvement in anaphora resolution lead to a marked improvement of NLP applications? (3)



Study 2: The impact of coreference resolution on NLP applications

- BART coreference resolution system
- Investigating the impact on:
  - Text summarisation
  - Text classification
  - Textual entailment

Text summarisation



The summarisation experiment

- Information from coreference resolver is used to increase score of each sentence by
  - Setting 1: score of longest mention in chain
  - Setting 2: highest score of mention in chain

  for each coreferential chain traversing the sentence

- Chains with one element (singletons) discarded

- Score of words calculated using their frequency in document without any morphological processing and with the stopwords filtered

The summarisation experiment (II)

- Corpus:
  - 89 randomly selected texts from the CAST corpus (http://clg.wlv.ac.uk/projects/CAST/corpus/)
  - Each text annotated with information about the importance of each sentence:
    - 15% marked as ESSENTIAL
    - a further 15% marked as IMPORTANT
- Evaluation:
  - Precision, recall, f-measure
  - Produced summaries of 15% and 30% compression rate

Results and discussion
summarisation experiment

Compression rate	15%	30%
Without BART	32.88%	46.34%
With BART – setting 1	28.62%	45.88%
With BART – setting 2	27.14%	45.19%

- Performance of summarisation decreases when coreference information is added
- Drop is less for 30% summaries
- Decrease in performance can be explained by the errors introduced by the coreference resolver

Text categorisation

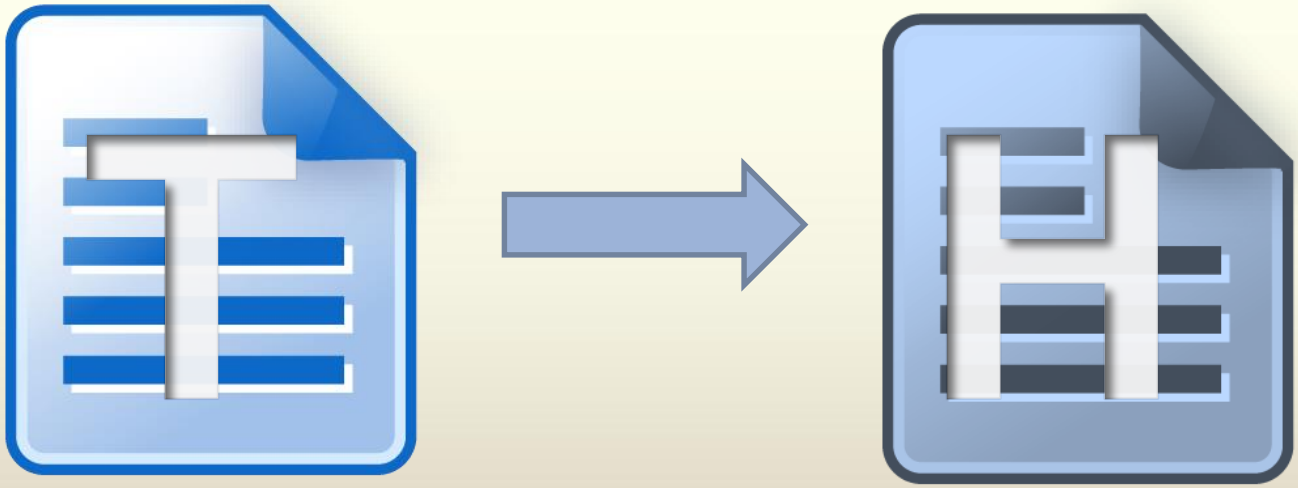
Results and Discussion

Text classification experiments

	P	R	F1
run-bow	95.59%	60.89%	74.39%
run-bart	95.70%	61.05%	74.54%

- Boosting *tfidf* weights of terms occurring in coreference chains **does not** significantly improve text classification performance
- Approach limitations:
  - Limited BART performance -> coreference information is noisy
  - BART biased towards named entities -> coreference chains are incomplete; common nouns could be more important
  - Feature selection -> could discard boosted terms
  - Results are quite high (95% macro averaged precision); perhaps a more challenging classification task would benefit more from coreference information

Textual entailment



Textual entailment experiments

- Classifier is trained on **similarity metrics**
  - Lexical similarity metrics (e.g. Precision, Recall)
  - BLEU (Papineni et al., 2002)
  - METEOR (Denkowski and Lavie, 2011)
  - TINE (Rios et al., 2011)
- Coreference chains processed: each mention in a chain is substituted by the longest (**most informative**) mention (Castillo 2010)
- Train/Test RTE two-way benchmark datasets

Results

Textual entailment experiments

- Accuracy with 10-fold-cross validation
- Comparison: model with coreference information and model without coreference information

Dataset	Model coref	Model no-coref
RTE-1	54.14	56.61
RTE-2	58.50	60
RTE-3	60.25	67.25

Results
Textual entailment experiments (2)

- Accuracy with test datasets
- Comparison: model with coreference information and model without coreference information

Dataset	Model coref	Model no-coref
RTE-1	56.87	56.87
RTE-2	57.12	59.12
RTE-3	60.25	61.75

Discussion

- For coreference resolution, impact of BART investigated
- BART has no positive impact
- Alternative models for coreference resolution should be considered as well
- Not-so-high performing anaphora or coreference resolution is not an encouraging option

Ways forward?

- Development of customised and domain-specific anaphora/resolution systems.
- Exploiting semantic knowledge (see also Soraluze et al.'s presentation at this workshop)
- Better pre-processing?
- Producing (*and sharing*) more resources.

Anaphora and coreference: 3 perennial questions

1. Are (automatic) anaphora resolution and coreference resolution beneficial to NLP applications?
2. Do we know how to evaluate anaphora resolution algorithms?
3. Which are the coreferential links most difficult to resolve?

The mystery of the original results

Intrinsic evaluation results

- MARS: success rate 45-65%
- Over this data: 46.63% (MARS'02), 49.47% (MARS'06)
- Our study of knowledge-poor approaches and full-parser approaches on 2,597 anaphors and 3 genres (Mitkov and Hallett 2007):
  - MARS: 57.03%
  - Kennedy and Boguraev: 52.08%
  - Baldwin's CogNIAC: 37.66%
  - Hobbs' naïve algorithm: 60.07%
  - Lappin and Leass RAP: 60.65%
  - Baselines: 30.07%-14.56%

The mystery of the original results

- Differences between results presented in the original papers and the results obtained in our study
- Hobbs (1976): 31.63%
- Lappin and Leass (1998): 25.35%
- Boguraev and Kennedy (1996): 22.92%
- Mitkov (1996, 1998): 31.97%
- Baldwin (1997): 54.34%

Why are results so different?

- Different genres (computer science manuals: ill-structured)
- Procedure fully automatic
- Lack of domain-specific NER

The issue of complexity of evaluation data

- Some evaluation data may contain anaphors which are more difficult to resolve such as
  - anaphors that are ambiguous and require real-world knowledge
  - anaphors that have a high number of competing candidates
  - anaphors that have their antecedents far away
- Other data may have most of their anaphors with single candidates for antecedent ⇒
- Resolution complexity has to be quantified for every evaluation data

Quantifying the complexity via the evaluation workbench

- Average referential distance in NPs between the anaphor and its antecedent (for each sample or all anaphors)
- Average referential distance in sentences between the anaphor and its antecedent (for each sample or all anaphors).

Difficult anaphors?

If Peter Mandelson had been in Tony Blair's shoes **he** would have demanded **his** resignation the day the Prime Minister forced **him** to leave the Cabinet.

Mysteries in evaluation

No sufficient evaluation details

Not clear what is the degree of automation of the system

Transparency, honesty?

Objectivity?

- How objective is evaluation?
- How objective are (annotated) corpora?
- How objective/reliable is human judgement?
- Interannotator agreement can be as low as 60% (Mitkov et al. 2000)

Reluctance...

- ... to publish modest or negative results
- Publishing negative results is also worthwhile!

Anaphora and coreference: 3 perennial questions

1. Are (automatic) anaphora resolution and coreference resolution beneficial to NLP applications?
2. Do we know how to evaluate anaphora resolution algorithms?
3. Which are the coreferential links most difficult to resolve?

Effects of Identity Degree of Anaphoric Relations on the Cognitive Effort of Readers (1)

- Research question 1: Does the degree of near-identity relations have an effect on the cognitive effort of readers who try to identify the antecedent of a specific anaphor?

- Data: Pairs of sentences from Recasens, Marti and Orasan (2012) with human annotation of **weak near identity** (class 1), **strong near identity** (class 2) and **total identity** (class 3).

- Statistical analysis: Eye tracking data from a preliminary study detected statistically significant differences between cases with identity degree 1 (weak identity) and 3 (total identity) in:
  - the time viewed measure ($p = 0.001$)
  - the number of gaze fixations measure ($p = 0.000$)

- Conclusion: The degree of identity of elements in a coreference chain affects the amount of cognitive effort required by readers to identify them as being coreferential

Effects of Identity Degree of Anaphoric Relations on the Cognitive Effort of Readers (2)

- Research question 2: Does the degree of identity relation have an effect on the cognitive effort of readers in cases where both the antecedent and the anaphor are **definite** noun phrases?

- Data: Selected snippets where both the antecedent and the anaphor were definite noun phrases (as opposed to indefinite ones).

- Statistical analysis: Statistically significant differences between cases with identity degree 1 (weak identity) and 3 (total identity) in:
  - the time viewed measure (p = 0.006)
  - the number of gaze fixations measure (p = 0.007)

- Conclusion: The degree of identity of elements in a coreference chain affects the amount of cognitive effort required by readers to identify them as being coreferential, regardless of whether or not they are both definite noun phrases.

Thank you very much

- Contact details
- My email: R.Mitkov@wlv.ac.uk
- My webpage: www.wlv.ac.uk/~le1825
- My research group web page: clg.wlv.ac.uk

Anaphora and coreference resolution: can they help NLP applications?

Ruslan Mitkov

with contributions from Richard Evans, Constantin Orăsan, Iustin Dornescu and Miguel Rios