Dutch Coreference Resolution: Issues and Applications

Veronique Hoste

LT3 Language and Translation Technology Team
Ghent University Association
http://veto.hogent.be/lt3

November 14, 2008

Background

As an alternative to knowledge-based approaches, corpus-based machine learning techniques have become increasingly popular for the resolution of coreferential relations.

Machine learning of coreference resolution

- Unsupervised: clustering task, combining noun phrases into equivalence classes.
  e.g. Cardie and Wagstaff, 99

Machine learning of coreference resolution

- Unsupervised: clustering task, combining noun phrases into equivalence classes.
  e.g. Cardie and Wagstaff, 99
- Supervised: requires an annotated corpus. Given two entities in a text, NP1 and NP2, classify the pair as coreferential or not coreferential. => coreference resolution as classification task.
  e.g. Aone and Bennett (1995), McCarthy (1996), Soon et al. (2001), Ng and Cardie (2002), and many others.

Typical supervised architecture

- Classify NP1 and NP2 as coreferential or not. The pair of NPs is represented by a feature vector containing distance, morphological, lexical, syntactic and semantic information on the candidate anaphor, its candidate antecedent and also on the relation between both.

Typical supervised architecture

- Classify NP1 and NP2 as coreferential or not. The pair of NPs is represented by a feature vector containing distance, morphological, lexical, syntactic and semantic information on the candidate anaphor, its candidate antecedent and also on the relation between both.
- In a postprocessing phase, a complete coreference chain has to be built between the pairs of NPs that were classified as being coreferential.

Annotation

Sources

MUC-7 manual, manual from Davies et al. (1998), critical remarks from Kibble (2000) and van Deemter and Kibble (2000).

Relations

- Identity relations between noun phrases, where both noun phrases refer to the same extra-linguistic entity.
- Bound relations where an anaphor refers to a quantified antecedent
- Predicative relations
- Super set–subset or group–member relations
  e.g. In the council meeting the confidence in [mayor-and-aldermen]$_1$ has been withdrawn. A motion requests that [all aldermen]$_2$ resign.

Annotation

Ongeveer een maand geleden stuurde < COREF ID = "1" > American Airlines < /COREF > < COREF ID = "2" MIN = "toplui" > enkele toplui < /COREF > naar Brussel. < COREF ID = "3" TYPE = "IDENT" REF = "1" MIN="vliegtuigmaatschappij" > De grote vliegtuigmaatschappij < /COREF > had interesse voor DAT en wou daarover < COREF ID = "5" > de eerste minister < /COREF > spreken. Maar < COREF ID = "6" TYPE = "IDENT" REF = "5" > Guy Verhofstadt < /COREF > (VLD) weigerde < COREF ID = "7" TYPE = "BOUND" REF = "2" > de delegatie < /COREF > te ontvangen.

Annotated material

Corpus	#docs	#tokens	#ident	#bridge	#pred	#bound
KNACK	267	122,960	9,179	na	na	43
DCOI	99	33,232	965	126	50	6
CGN	29	20,812	2,077	296	147	15
IMIX	497	135,828	4,910	1,772	289	19

Inter-annotator agreement

- 29 documents from CGN and DCOI; 2 annotators
- For the ident relation: inter-annotator agreement as the F-measure of the MUC-scores obtained by taking one annotation as 'gold standard' and the other as 'system output'.
- For the other relations: inter-annotator agreement as the average of the percentage of *anaphor-antecedent* relations in the gold standard for which an *anaphor-antecedent′* pair exists in the system output, and where *antecedent* and *antecedent′* belong to the same cluster (w.r.t. the IDENT relation) in the gold standard.
- Agreement:
  - IDENT: 76%
  - BRIDGING: 33%
  - PRED: 56%
  - No agreement on the (small number of) BOUND relations.

Main sources of disagreement

- Cases where an annotator fails to annotate a coreference relation.
- Cases where a BRIDGE or PRED relation is annotated as IDENT.
- Cases where multiple interpretations are possible.
- Unclear guidelines. It was unclear whether titles and other leading material from news items should be considered part of the annotation task. It was unclear which appositions should be annotated with a PRED relation.

Instance construction

- Per NP type (Pronouns/Proper nouns/Common nouns)
- Positive: anaphor + each preceding element in the chain
- Negative: anaphor + each preceding NP not in the chain (search scope: $<=$ 20 sentences)
- Highly skewed class distribution:
  positive: 6,457 inst. (KNACK-2002)
  negative: 95,919 inst. (KNACK-2002)

Instance construction

- Positional features (eg. dist_sent, dist_NP)
- Local context features
- Morphological and lexical features (e.g. i/j/ij-pron, j_demon, j_def, i/j/ij-proper, num_agree)
- Syntactic features (e.g. i/j/ij_SBJ/OBJ/PREDC, appositive)
- String-matching features (comp_match, part_match, alias, same_head)
- Semantic features (synonym, hypernym, same_NE, (linguistic) gender of antecedent and anaphor, semantic class of NP)

Additional semantic information

- Unsupervised k-means clustering on Dutch news corpus: top-10,000 nouns/names clustered into 1000 groups based on the similarity of their syntactic relations (Van de Cruys, 2005)
- e.g. 201 barrière belemmering drempel hindernis hobbel horde knelpunt obstakel struikelblok
  (English: barrier impediment threshold hindrance bump hurdle bottleneck obstacle block)
- Presence of noun in a cluster represented in 3 Features: clust_anaphor, cluster_antecedent, same_clust
- Related work: Ji et al. (2005), Ng (2007), Ponzetto and Strube (2006)

Additional syntactic information

- Produced by the Alpino parser (Bouma, 2001)
- Additional features:
  - Dependency label as predicted for (the head word of) the anaphor and for the antecedent.
  - Dependency path between the governing verb and the anaphor, and between the verb and antecedent.
  - Clause information: is the anaphor / antecedent part of the main clause or not.
  - Root Overlap: binary feature that codes overlap between 'roots' or lemmas of the anaphor and antecedent.
- Related work: Luo and Zitouni (2005), Yang et al. (2006)

Additional syntactic information

Example

Algemeen directeur Jan Gijsen van Ford Genk maakt bekend dat het bedrijf de volgende twee jaar 1400 banen wil schrappen. (English: Head director Jan Gijsen of Ford Genk announces that the company will cut 1400 jobs in the next two years.)

dependency label anaphor: subject
dependency label antecedent: object1
label match: no
dependency path anaphor: [[schrap,hd/su],[wil,hd/su]]
dependency path antecedent:
med[[maak_bekend,hd/su,directeur,hd/mod,van,hd/obj1]]
clause anaphor: not in main clause
clause antecedent: in main clause
root overlap: no

Issues

- Error percolation
- Lack of semantic resources
- Two-step classification approach
- Evaluation

Which ML approach?

Background

- Each learner has a different bias
  = the search heuristics a certain machine learning method uses and the way it represents the learned knowledge
  E.g. decision tree learners favor compact decision trees

Which ML approach?

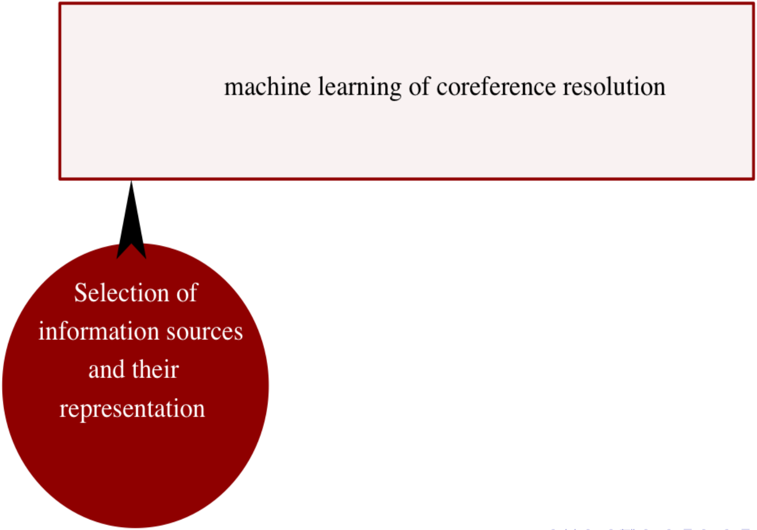
Background

- Each learner has a different bias
  = the search heuristics a certain machine learning method uses and the way it represents the learned knowledge
  E.g. decision tree learners favor compact decision trees
- No free lunch theorem (Wolpert and Macready, 1995)
  = no inductive algorithm is universally better than any other

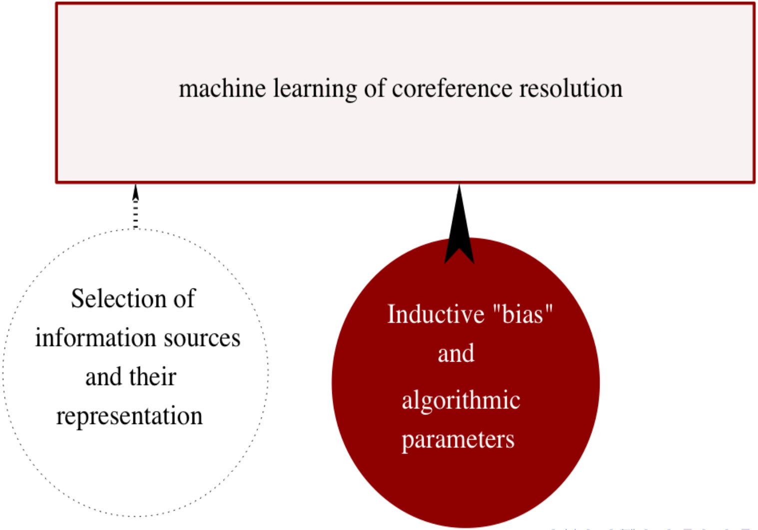
Typical ML architecture

2 or more algorithms are compared for a fixed sample selection, feature selection and representation over a number of trials. Sometimes learning curves, limited parameter optimization.

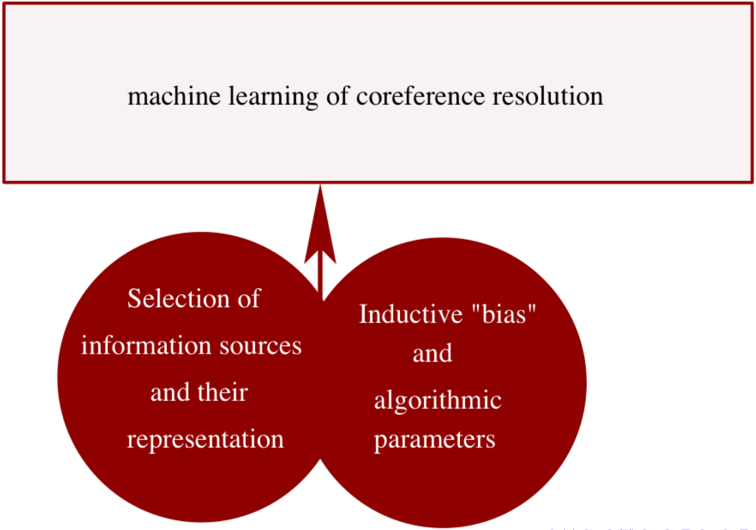
What influences the outcome of a ML experiment?



What influences the outcome of a ML experiment?



What influences the outcome of a ML experiment?



Experiment

Investigate the effect of

- feature selection (e.g. backward selection)
- algorithm parameter optimization
- sample selection
- interleaved feature selection and parameter optimization

on the comparison of two inductive algorithms (lazy and eager) on the task of coreference resolution

Lazy versus eager

Memory-based learning (MBL)

- performance in real-world tasks is based on remembering past events rather than creating rules or generalizations
- Lazy: MBL keeps all training data in memory and at classification time, the similarity of an unseen test item to all examples in memory is computed using a similarity metric. The class of the most similar example(s) is then used as prediction for the test instance.
- TiMBL (Daelemans et al., 2002)

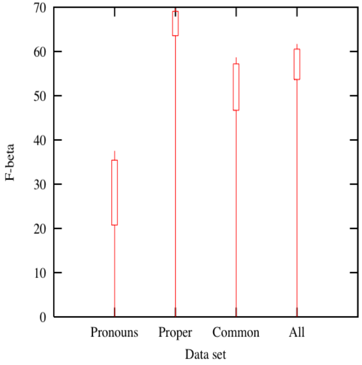
Rule induction

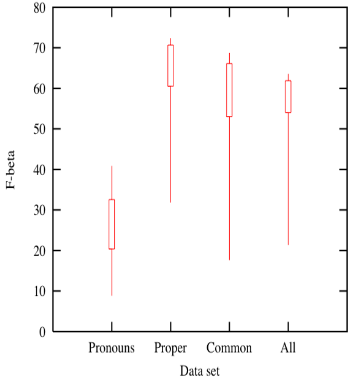
- Minimal-description-length-driven or eager: compress the training material by extracting a limited number of rules.
- Ripper (Cohen, 1995)

Feature selection

	TIMBL				RIPPER			
All	Acc.	Prec.	Rec.	$F_{\beta=1}$	Acc.	Prec.	Rec.	$F_{\beta=1}$
default	94.29	56.80	55.50	56.15	96.09	84.65	49.65	62.59
backward	95.73	76.38	50.98	**64.14**	96.12	84.98	49.98	**62.94**
GR	95.58	81.09	42.86	56.08	95.58	81.09	42.86	56.08
bi.hill.	95.93	77.88	53.41	63.36	95.75	79.77	47.51	59.55
PPC								
default	94.35	57.19	56.21	56.70	95.98	79.73	52.59	63.16
backward	95.42	67.24	59.33	**63.04**	96.19	82.88	53.17	**64.78**
GR	95.71	88.89	39.85	55.03	95.72	89.59	39.55	54.88
bi.hill.	96.05	84.75	48.84	61.97	96.31	88.29	50.68	64.40
Pronouns								
default	91.88	38.33	27.42	31.97	93.27	54.78	19.44	28.70
backward	92.31	43.53	35.24	**38.95**	93.57	59.25	24.43	**34.59**
GR	93.04	0.00	0.00	0.00	93.04	0.00	0.00	0.00
bi.hill.	93.68	60.86	25.97	36.41	93.86	77.19	16.70	27.46
Proper nouns								
default	94.34	63.34	67.53	65.37	96.02	83.89	61.60	71.04
backward	94.97	67.86	69.34	68.59	96.13	86.10	60.90	71.34
GR	95.97	89.46	55.65	68.62	95.98	90.22	55.19	68.49
bi.hill.	96.26	89.67	59.57	**71.58**	96.28	90.17	59.52	**71.70**
Common Nouns								
default	95.41	53.70	53.53	53.62	97.09	79.61	55.55	65.44
backward	97.23	82.42	56.12	**66.77**	97.38	85.62	56.74	**68.25**
GR	96.56	87.38	35.87	50.87	96.57	87.90	35.70	50.77
bi.hill.	96.84	85.43	43.64	57.77	97.39	87.14	55.46	67.78

Parameter optimization





Conclusions

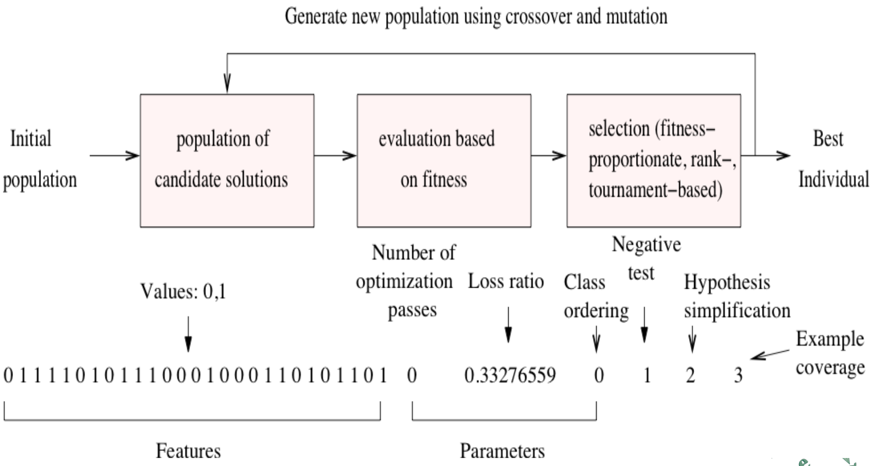
Feature selection

- Large effect of feature selection on classifier performance
- Especially TIMBL seemed to be very sensitive to a good feature subset
- The feature selection considered to be optimal for TIMBL could be different from the one optimal for RIPPER

Parameter optimization

The vertical performance differences are much larger than the horizontal algorithm-comparing performance differences.

Joint feature selection and parameter optimization



Joint feature selection and parameter optimization

KNACK-2002	DEFAULT			GA OPTIMIZATION		
TIMBL	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$
All	48.78	44.93	46.78	**71.83**	**45.50**	**55.71**
PPC	49.75	44.90	47.20	**70.22**	**49.74**	**58.24**
Pronouns	50.11	44.81	47.31	**67.65**	**53.04**	**59.46**
Proper nouns	62.84	54.04	58.11	**80.07**	**54.87**	**65.11**
Common nouns	30.65	30.37	30.51	**59.58**	**33.49**	**42.88**
RIPPER	Prec.	Rec.	$F_{\beta=1}$	Prec.	Rec.	$F_{\beta=1}$
All	**69.49**	34.92	46.49	61.51	**61.93**	**61.72**
PPC	**66.34**	41.75	51.25	60.68	**62.26**	**61.46**
Pronouns	**61.08**	43.14	50.57	58.95	**69.69**	**63.87**
Proper nouns	**76.84**	49.49	60.21	69.36	**62.71**	**65.87**
Common nouns	**61.82**	25.92	36.52	51.57	**43.48**	**47.18**

The problem of imbalanced data sets

As a consequence of recasting the problem as a classification task, coreference resolution data sets reveal large class imbalances: only a small part of the possible relations between noun phrases (NPs) is coreferential.

Filters

Goal

In order to cope with these class imbalances, different instance selection techniques have been proposed to rebalance the corpus

Goal: produce better performing classifiers

Procedure: filters split the basic set of instances in two parts: one parts gets a label automatically assigned by the filter, the other part is classified by a classifier.

Filters

Two different perspectives

- A language engineering approach, a preprocessing trick

Filters

Two different perspectives

- A language engineering approach, a preprocessing trick
- A principled approach to creating hybrid knowledge-based and machine learning based systems where both approaches solve the problems they are best at.

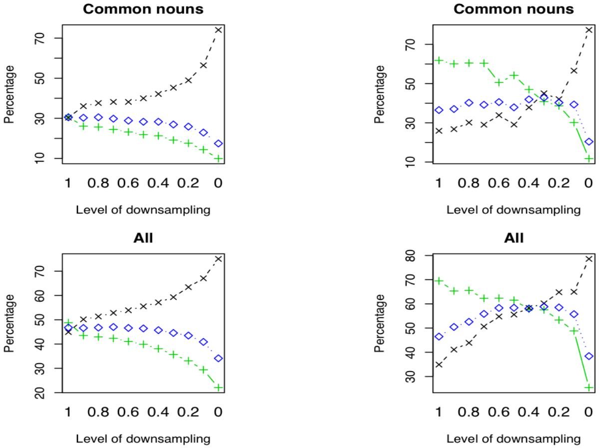
Random downsampling

- Rebalancing of the data is done without any a priori knowledge about the task to be solved and linked to the specific learning behaviour of a lazy learner (TIMBL) and an eager learner (RIPPER) (Hoste 2005)

Random downsampling

- Rebalancing of the data is done without any a priori knowledge about the task to be solved and linked to the specific learning behaviour of a lazy learner (TIMBL) and an eager learner (RIPPER) (Hoste 2005)
- Learning approaches can behave quite differently in case of skewness of the classes and they also react differently to a change in class distribution.

Effect of random downsampling



Linguistically motivated filters

e.g. Strube et al. 2002,Yang et al. 2003, Ng and Cardie 2002, Harabagiu et al. 2001,Uryupina 2004.

- Negative sample selection: filters aiming at the reduction of negative instances, reducing the positive class skewness.
  e.g. Strube et al. 2002: reduction of 50% of the negative instances.
  e.g. discard an antecedent-anaphor pair if the anaphor is an indefinite NP

Linguistically motivated filters

e.g. Strube et al. 2002,Yang et al. 2003, Ng and Cardie 2002, Harabagiu et al. 2001,Uryupina 2004.

- Negative sample selection: filters aiming at the reduction of negative instances, reducing the positive class skewness.
  e.g. Strube et al. 2002: reduction of 50% of the negative instances.
  e.g. discard an antecedent-anaphor pair if the anaphor is an indefinite NP
- Negative and positive sample selection: one antecedent is sufficient to resolve an anaphor. e.g. Ng and Cardie 2002, Harabagiu et al. 2001

Effect of the different possible filters

- fdef: filters out all instances containing indefinite anaphora
- The filter fhead filters out instances in which the anaphor and antecedent are located at a distance of more than three sentences from each other and do not share the same head word
- The filter fagree applies to pronouns only and demands agreement between anaphor and antecedent.
- The filter rule fmatch (cf. Ng and Cardie 2002) assigns a positive label to an instance that describes an anaphor and antecedent which have a complete string match.
- The filter f3s restricts the search space for pronouns to three sentences.

Cross validation experiments on the training set with and without the different filters

	MAXENT	TIMBL	#num.	MAXENT	TIMBL
default	37.6	46.7	76,920	37.6	46.7
fdef	37.6	44.2	64,656	40.0	46.8
fagree	37.9	44.7	66,786	39.5	46.4
f3s	31.6	35.2	59,183	41.5	45.2
fhead	34.8	39.7	15,041	58.3	67.0
fmatch	43.1	43.6	57,479	39.0	39.7
combi1	29.3	31.3	9,723	65.9	70.8
combi2	31.5	30.5	6,286	55.6	54.0

MUC scores on test set with and without the different filters

	TIMBL			MAXENT		
	recall	precision	F-score	recall	precision	F-score
normal	**60.0**	**35.2**	**44.4**	**41.7**	**42.2**	**42.0**
fdef	49.2	46.7	47.9	39.5	46.4	42.7
f3s	58.0	36.8	45.1	51.2	43.8	47.2
fagree	50.2	40.4	44.7	41.3	42.3	41.8
fhead	39.8	60.3	47.9	45.5	42.7	44.1
fmatch	46.7	48.4	47.5	51.2	42.4	46.4
combi1	40.7	46.1	43.2	38.5	51.6	44.1
combi2	**36.7**	**61.0**	**45.8**	**40.0**	**51.8**	**45.1**

all hybrid systems: improving the precision of the system at the expense of recall

One-class learning

- The common approach in detection tasks is to define these tasks as two-class classification problems: the classifier labels instances as being "coreferential" or "non-coreferential"
- But why not consider it as one-class classification? (e.g. Manevitz, 2001)

One-class learning

Motivation

We are only given examples of one class, namely of coreferential relations between NPs and we wish to determine whether a pair of NPs is coreferential. But the negative "non-coreferential" class can be anything else, which makes the choice of negative data for this task arbitrary.

One-class learning

Experiment

- One-class SVM's on the positive examples in the training set
- Results:

	Prec.	Rec.	F-score
default (rbf kernel)	39.9%	58.6%	47.5%
one-class (rbf kernel)	74.9%	28.4%	41.2%

- Compact, dense region in the example space?

Applications

- Information Extraction module for the medical domain
- Question-Answering

Information Extraction module for the medical domain

Application

- construct a Relation Finder which can predict medical semantic relations.
- corpus: version of the Spectrum medical encyclopedia in which sentences and noun phrases are annotated with domain specific semantic tags.

Examples

<rel_treats id="19"> Veel gevallen van <con_disease id="6"> asfyxie</con_disease> kunnen door <con_treatment id="14"> beademing </con_treatment>, of door opheffen van de passagestoornis (<con_treatment id="15"> tracheotomie </con_treatment>) weer herstellen. </rel_treats>

Information Extraction module for the medical domain

Experiment

- Relation finder: a maximum entropy modeling algorithm trained on approximately 2000 annotated entries of MedEnc. (avg. 10 sent)
- Two separate test sets of 50 and 500 entries respectively
- Two experiments: one using the predicted coreference relations as features, and one without these features.

test set	without	with
small(50)	53.03	53.51
Big(500)	59.15	59.60

Question-Answering

Experiment

- Similar information extraction experiment, concentrating on relations where at least one of the arguments is a named entity, such as date-of-birth, age, capital-of, and founder-of.

Question-Answering

Experiment

- Similar information extraction experiment, concentrating on relations where at least one of the arguments is a named entity, such as date-of-birth, age, capital-of, and founder-of.
- After adding coreference resolution, the number of extracted facts goes up with over 50% (from 93K to 145K).

Question-Answering

Experiment

- Similar information extraction experiment, concentrating on relations where at least one of the arguments is a named entity, such as date-of-birth, age, capital-of, and founder-of.
- After adding coreference resolution, the number of extracted facts goes up with over 50% (from 93K to 145K).
- Incorporation of these facts into a Question Answering system leads to an improvement in accuracy of 5% (from 65% to 70%) on questions of the appropriate type.

Future work

DuOMAn project

- Sentiment detection in Dutch blogs
- Cross-document coreference resolution

SoNaR project

- Multi-level annotation project

Thank you!