Anaphora Resolution:

Theory and Practice

Michael Strube
European Media Laboratory GmbH
Heidelberg, Germany
Michael.Strube@eml.villa-bosch.de

Anaphora Resolution:

Theory ~~and~~ Practice

Michael Strube

European Media Laboratory GmbH

Heidelberg, Germany

Michael.Strube@eml.villa-bosch.de

Anaphora Resolution:

or

Theory ~~and~~ Practice

Michael Strube

European Media Laboratory GmbH

Heidelberg, Germany

Michael.Strube@eml.villa-bosch.de

A Few Questions

- How do insights taken from centering-based models fare if they are applied to large amounts of naturally occurring data?

- How do centering-based models compare to corpus-based methods? Results? Coverage? Portability? Robustness? Development time?

- What do centering-based models and corpus-based methods have in common? What are the differences?

A Few Questions

- How do insights taken from centering-based models fare if they are applied to large amounts of naturally occurring data?
- How do linguistic theories fare if they are applied to large amounts of naturally occurring data?
- How do centering-based models compare to corpus-based methods? Results? Coverage? Portability? Robustness? Development time?
- How do linguistic theories compare to corpus-based methods? Results? Coverage? Portability? Robustness? Development time?
- What do centering-based models and corpus-based methods have in common? What are the differences?

$\Rightarrow$ What is our linguistic intuition good for?

Overview

1. look back at *Never look back*;
2. NLB applied to spoken dialogue;
3. machine learning approach to reference resolution in text (how much annotated data is needed to train an anaphora resolution classifier?);
4. machine learning approach to pronoun resolution in spoken dialogue (which features do the work?);
5. concluding remarks.

Never Look Back: An Alternative to Centering (NLB)

Motivation:

- centering accounts for intra-sentential anaphora by means of an appropriate definition of the *utterance*;
- however, the *utterance*, which is the most crucial element in centering, is not specified in the original literature (e.g. Grosz et al. (1995));
- Kameyama (1998) presented an elaborate model on *intra-sentential* centering; however, that model still cannot be applied to unrestricted data;
- Kehler (1997) observed that centering is not cognitively plausible due to it's lack of incrementality.

NLB: The Model

- (Discard most of the centering machinery.)
- One construct: The list of salient discourse entities (S-list).
- Two operations on the S-list:
  1. Incremental update: Insertion of discourse entities;
  2. Periodic elimination of discourse entites: Removing of discourse entities which are not realized in the immediately preceding elimination unit.
- S-list describes the attentional state of the hearer at *any* given point in processing a discourse.
- Order among elements of the S-list directly provides preferences for interpretation of pronouns.

NLB: The Algorithm

1. If a referring expression is encountered,

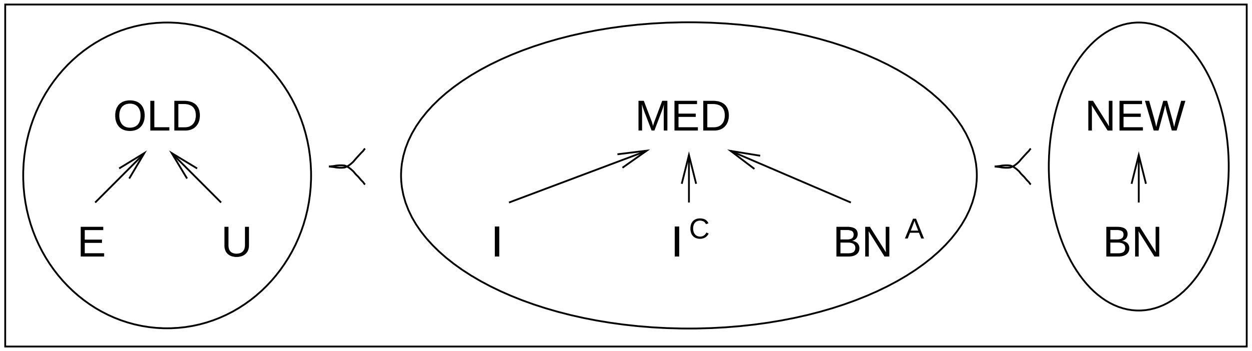
   (a) if it is a pronoun, test the elements of the S-list in the given order until the test succeeds;

   (b) update S-list; the position of the discourse entity associated with the referring expression under consideration is determined by the S-list-ranking criteria which are used as an insertion algorithm.

2. If the analysis of elimination unit *U* is finished, remove all discourse entities from the S-list, which are not *realized* in *U*.

NLB: The S-list Ranking

- familiarity:



- linear order.

NLB: Results

- results obtained by hand-simulation of the algorithm;
- two languages: English and German (for each language about 600 pronouns);
- about 10% improvement in success rate over previous centering-based approaches (results confirmed by Tetrault (2001) who implemented a simplified version and compared that with a syntax-based version, which did even better).

NLB: Conclusions

- pronoun resolution requires incremental update of the discourse representation and incremental resolution (I consider Tetrault’s (2001) results as confirmation of that point);
- the incremental update helps to deal with pronouns with intra- and intersentential antecedents;
- there is no need for centering constructs like *backward-looking center*, *forward-looking centers* and *centering transitions*;
- (orthodox) centering may be a help for a lot of tasks in NLP, but definitely not for pronoun resolution.

Application: Anaphora Resolution in Spoken Dialogue

(Joint work with Miriam Eckert, formerly at UPenn, now at Microsoft)

Spoken Dialogue is Messy!

```
B.57:  -- what are they actually telling us, and after,
       you know, what happened the other day with that,
       uh, C I A guy, you know --
A.58:  Uh-huh.
B.59:  -- how much is, what all the wars we’re getting
       into and all the, you know, the messes we’re --
A.60:  That we really don’t --
B.61:  -- we’re bombing us, ourselves with --
A.62:  -- that we don’t know about,
B.63:  -- right, is that true, or, you know, is it,
A.64:  How much,
B.65:  is (( )),
A.66:  of it’s true, and how much --
B.67:  really a threat,
A.68:  -- and how much of it is propaganda --
B.69:  Right.
       (sw3241)
```

Anaphora Resolution in Spoken Dialogue: Problems

- center of attention in multi-party discourse;
- utterances with no discourse entities;
- abandoned or partial utterances (disfluencies, hesitations, interruptions, corrections);
- determination of *utterance units* (no punctuation in spoken dialogue!);
- low frequency of individual anaphora (NP-antecedents: 45.1%), but high frequency of discourse-deictic (non-NP-antecedents: 22.6%) and vague (no antecedents: 32.3%) anaphora (data based on only three Switchboard dialogues).

Types of Anaphora I: Individual – 45.1%

(IPro, IDem)

(4) **A:** **He**$_i$[McGyver]'s always going out and inventing new things out of scrap [...]

B: Boeing ought to hire **him**$_i$ and give **him**$_i$ a junkyard$_j$, . . . and see if **he**$_i$ could build a Seven Forty-Seven out of **it**$_j$.

(sw2102)

Types of Anaphora II: Discourse-Deictic – 22.6%

(DDPro, DDDem)

(5) **A:** [The government don't tell you everything.]$_i$
B: I know **it**$_i$.
(sw3241)

(6) **A:** ...[we never know what they're thinking]$_i$.
B: **That**$_i$'s right. [I don't trust them]$_j$,
maybe I guess **it**$_j$'s because of what happened over there
with their own people, how they threw them out of power...
(sw3241)

Types of Anaphora III: Vague – 13.2%

(VagPro, VagDem)

(7) **B.27** She has a private baby-sitter.
A.28 Yeah.
B.29 And, uh, the baby just screams. I mean, the baby is like seventeen months and she just screams.
A.30 Uh-huh.
B.31 Well even if she knows that they're fixing to get ready to go over there. They're not even there yet –
A.32 Uh-huh.
B.33 – you know.
A.34 Yeah. **It**'s hard.

Types of Anaphora IV: Inferrable-Evoked Pronouns – 19.1%

(IEPPro)

(7) **A:** I think the **Soviet Union** knows what we have
and knows that we're pretty serious and if **they** ever tried
to do anything, we would, we would be on the offensive.
(sw3241)

Proposal for Pronoun Resolution in Spoken Dialogue I

1. use *update* and *elimination unit*, but redefine *elimination unit* in terms of dialogue acts (pairs of initiations and acknowledgements; acknowledgments signal that common ground is achieved);

2. classify different types of anaphora using the predicative context of the anaphor;

3. resolve individual and discourse-deictic anaphora.

Proposal for Pronoun Resolution in Spoken Dialogue II

Classification of different types of pronouns and demonstratives, so that

- resolution of individual anaphora is only triggered if anaphor is classified as individual ($\rightarrow$ *A-incompatible*);
- resolution of discourse-deictic anaphora is only triggered if anaphor is classified as discourse-deictic ($\rightarrow$ *I-incompatible*);

A-Incompatible (*A)

x is an anaphor and *cannot* refer to abstract entities.

- Equating constructions where a pronominal referent is equated with a concrete individual referent, e.g., *x is a car.*

- Copula constructions whose adjectives can only be applied to concrete entities, e.g., *x is expensive, x is tasty, x is loud.*

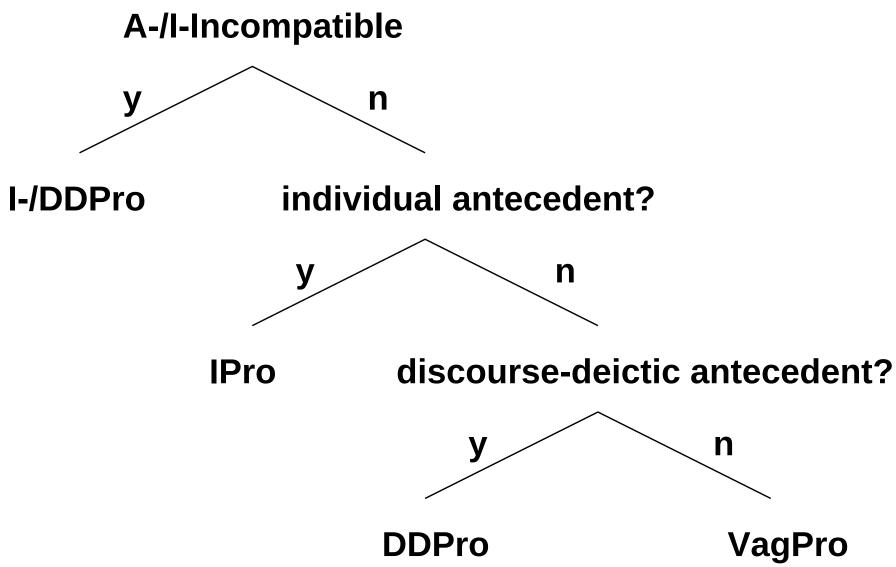
- Arguments of verbs describing physical contact/stimulation, which cannot be used metaphorically, e.g., *break x, smash x, eat x, drink x, smell x* but NOT **see x*

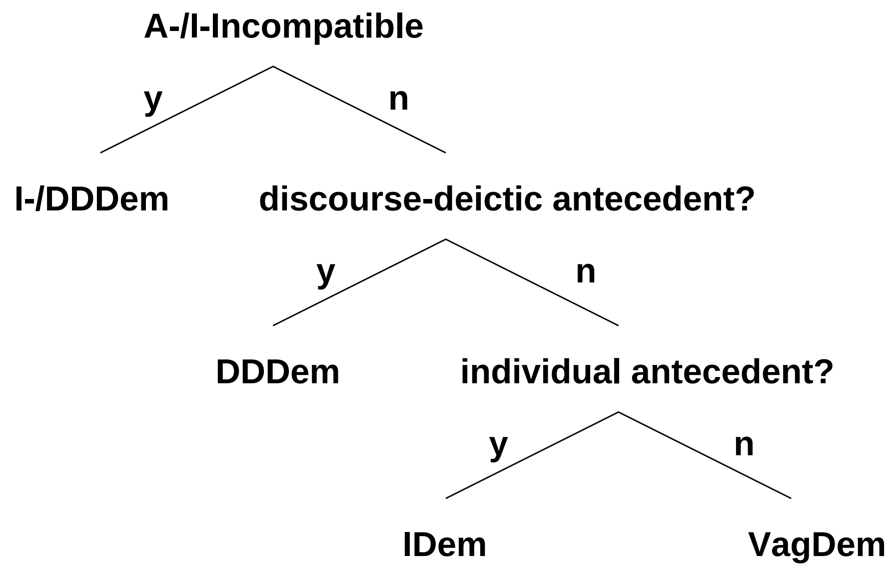
I-Incompatible (*I)

x is an anaphor and *cannot* refer to individual, concrete entities.

- Equating constructions where a pronominal referent is equated with an abstract object, e.g., *x is making it easy, x is a suggestion.*
- Copula constructions whose adjectives can only be applied to abstract entities, e.g., *x is true, x is false, x is correct, x is right, x isn't right.*
- Arguments of verbs describing propositional attitude which *only* take S'-complements, e.g., *assume.*
- Object of *do*.
- Predicate or anaphoric referent is a "reason", e.g., *x is because I like her, x is why he's late.*

Overview of the Algorithm





The Algorithm: Switchboard sw3117

28-	I	B.18	And **[she** ended up going to the **University of Oklahoma]**.
-28	A	A.19	Uh-huh.
			S: [DAUGHTER: ***she***, U. OF OKLA.: ***U. of Okla.***]
29-29	I	B.20	I can say **that** because **it** was a big well known school,
			S: [U. OF OKLA.: ***it***]
			A: [SHE ENDED UP . . . : ***that***]
30-30	I		**it** had **a well known education** –
			S: [U. OF OKLA.: ***it***, EDUCATION: ***education***]

Anaphora Resolution in Spoken Dialogue: Summary

- attempt to resolve anaphora in naturally occurring dialogue;
- instead of ignoring words like *uh-huh, yeah, . . .* , we actually use them for recognizing when common ground is achieved;
- treatment of interruptions, hesitations, etc.;
- achieve higher precision than ordinary anaphora resolution algorithms by classifying different types of anaphors (baseline for individual anaphora would be around 30%);
- results published at EACL ’99, Amstelogue ’99, Journal of Semantics (17(1)).

Problems with These Approaches

- only some of the features explicit; some of them hidden by the mechanism (algorithm); this is much worse with original centering;
- there is only one S-list ordering; difficult to apply to different phenomena (e.g. pronouns vs. defNPs);
- difficult to evaluate the contribution of each feature.

Proposal: Machine Learning for Anaphora Resolution

- by applying ML-techniques we are able to determine more variables;
- most features are explicit;
- it is easy (though time-consuming) to determine the contribution of each feature; thus it is possible to decide whether it is necessary to include *expensive* features.

Application: Anaphora Resolution in Written Text

(joint work with Stefan Rapp (Sony Research) and Christoph Müller (EML))

- ACL ’02: *Applying Co-Training to Reference Resolution*
- EMNLP ’02: *The Influence of Minimun Edit Distance on Reference Resolution*
- is it possible to apply Co-Training (a weakly supervised machine learning meta algorithm) to reference resolution?
- German corpus (in the meantime applied to English as well with similar results);
- Co-Training did not work out that well (Ng & Cardie (2003, NAACL) report better results);
- however, by doing the Co-Training experiments we got interesting secondary results (reported at ACL ’02 and EMNLP ’02).

Features

- NP-level features
- coreference-level features

NP-level Features

	Document level features	
1.	doc_id	document number (1 . . . 250)
	NP-level features	
2.	ante_gram_func	grammatical function of antecedent (subject, object, other)
3.	ante_npform	form of antecedent (definite NP, indefinite NP, personal pronoun, demonstrative pronoun, possessive pronoun, proper name)
4.	ante_agree	agreement in person, gender, number
5.	ante_semanticclass	semantic class of antecedent (human, concrete and abstract object)
6.	ana_gram_func	grammatical function of anaphor (subject, object, other)
7.	ana_npform	form of anaphor (definite NP, indefinite NP, personal pronoun, demonstrative pronoun, possessive pronoun, proper name)
8.	ana_agree	agreement in person, gender, number
9.	ana_semanticclass	semantic class of anaphor (human, concrete object, abstract object)

Coreference-level Features

	Coreference-level features	
10.	wdist	distance between anaphor and antecedent in words (1 . . . n)
11.	ddist	distance between anaphor and antecedent in sentences (0, 1, $>$1)
12.	mdist	distance between anaphor and antecedent in markables (1 . . . n)
13.	syn_par	anaphor and antecedent have the same grammatical function (yes, no)
14.	string_ident	anaphor and antecedent consist of identical strings (yes, no)
15.	substring_match	one string contains the other (yes, no)
	New coreference-level features	
16.	ante_med	minimum edit distance to anaphor $ante\_med = 100 \cdot \frac{m-(s+i+d)}{m}$
17.	ana_med	minimum edit distance to antecedent $ana\_med = 100 \cdot \frac{n-(s+i+d)}{n}$

Different Results for Different Types of Anaphora

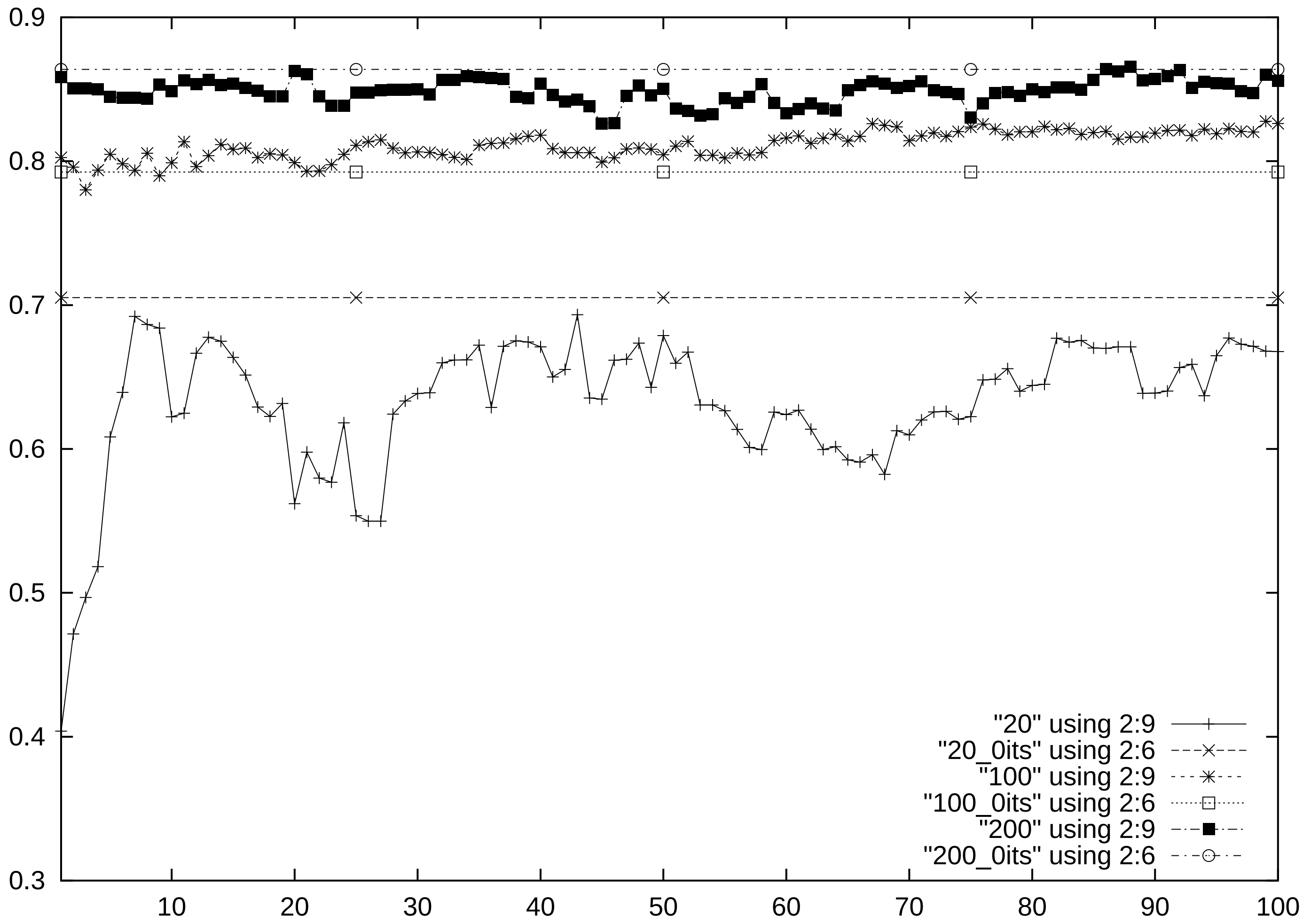
(from EMNLP ’02)

	P	R	F
defNP	69.26%	22.47%	33.94%
NE	90.77%	65.68%	76.22%
PDS	25.00%	11.11%	15.38%
PPER	85.81%	77.78%	81.60%
PPOS	82.11%	87.31%	84.63%
all	84.96%	56.65%	67.98%

Conclusions on Anaphora Resolution in Written Text

- it is useful to train a separate classifier for each NP form;
- we seem to be missing something with respect to defNP resolution: world/domain knowledge?
- resolution of proper names is ok;
- task of pronoun resolution in written text seems to be solved;
- from the Co-Training experiments, we conclude that the amount of training data required is surprisingly low.

Low Amount of Training Data for Pronouns



Amount of Training Data for Anaphora Resolution

- for training a pronoun resolution classifier only 100-200 labeled instances are needed;
- for training classifiers for resolving proper names and definite NPs rather 1000-2000 labeled instances are needed;
- this may be due the fact, that pronoun resolution classifiers only rely on a few simple features.

Application: Pronoun Resolution for Spoken Dialogue

(joint work with Christoph Müller (EML), ACL 2003)

- test whether insights taken from the Eckert & Strube algorithm for pronoun resolution in spoken dialogue work in a ML environment;
- port the system (environment) from German written text to English spoken dialogue;
- what are the results for different types of anaphoric expressions?
- determine which features contribute significantly to the results for each type of anaphoric expression.

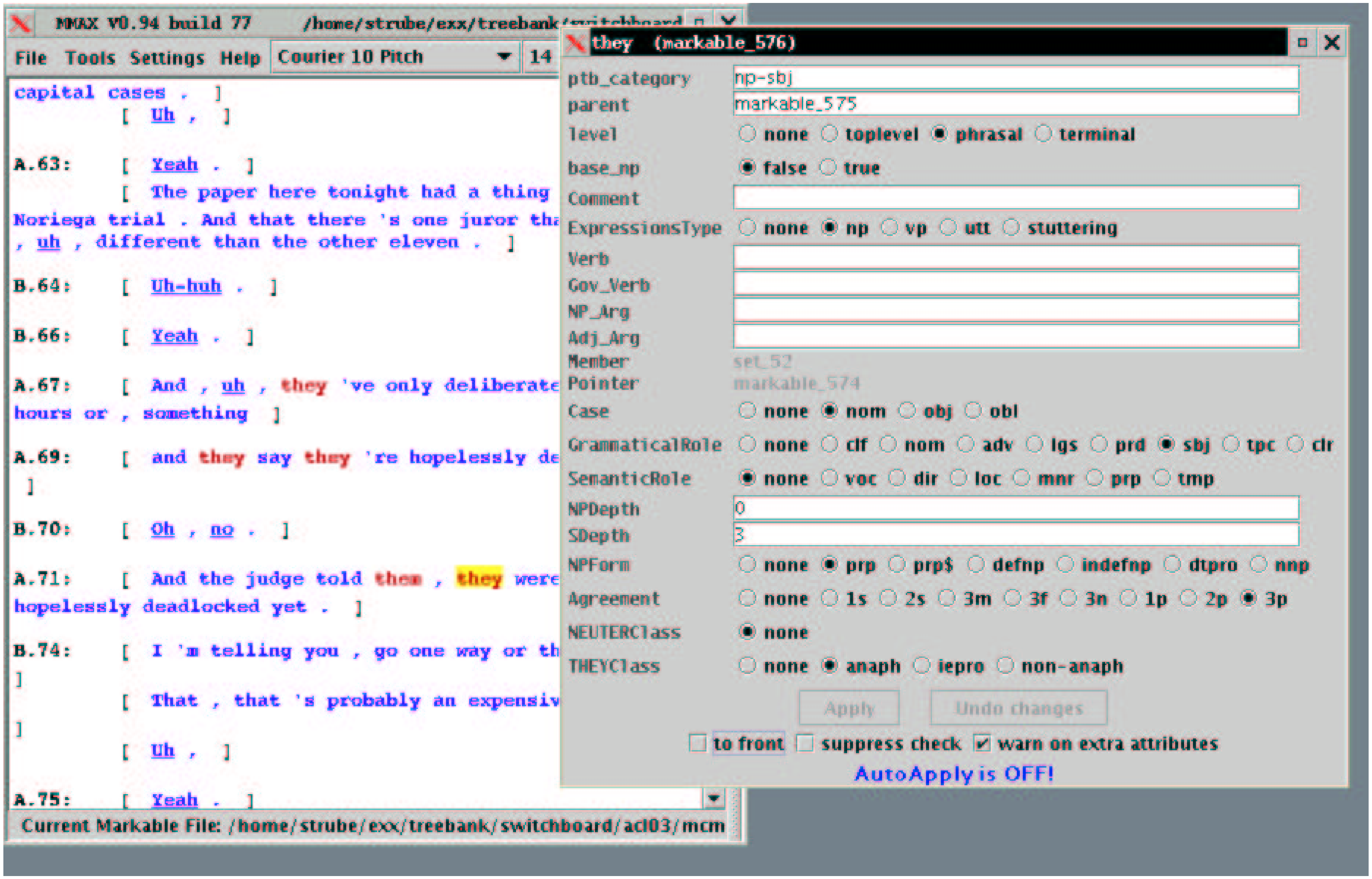
Hypotheses Derived from Previous Work

- pronoun resolution in written text works reasonably well without considering knowledge- or semantics-based features;
- for determining non-NP-antecedents of pronouns domain knowledge seems to be necessary (Byron, 2002);
- Eckert & Strube (2000) had the hypothesis that information about subcategorization frames of verbs gathered from corpora (e.g. Briscoe & Caroll (1997)) could be sufficient.

Data: Corpus

- 20 randomly chosen Switchboard dialogues;
- 30810 tokens (words, punctuation) in 3275 sentences/1771 turns;
- annotation consists of 16601 NP- and non-NP-markables;
- identification of markables and assignment of attributes guided by Penn Treebank;
- 924 third person neuter pronouns: 43.4% have NP-antecedents, 23.6% non-NP-antecedents, 33% have no antecedents (almost identical to the numbers reported by Eckert & Strube, 2000).

Data: Annotation



Data: Distribution of Agreement Features for Pronouns

	3m		3f		3n		3p	
prp	67	63	49	47	**541**	**318**	**418**	**358**
prp$	18	15	14	11	**3**	**3**	**35**	**27**
dtpro	0	0	0	0	**380**	**298**	**12**	**11**
$\sum$	85	78	63	58	**924**	**619**	**465**	**396**

- high number of singletons (223 for *it*, 60 for *they*, 82 for *that*);
- these are either expletive or vague and do not have antecedents marked in the corpus.

Data Generation for ML

- pronoun resolution viewed as binary classification;
- training and testing instances are pairs of potentially anaphoric pronouns and potential antecedents;
- instances are labeled *P* if both markables have the same value in their *member* attribute, *N* otherwise;
- pairs containing non-NP-antecedents restricted to cases where the pronoun was realized by *it* or *that* and the antecedent were non-NP-markables from the last two sentences.

Features

- NP-level features;
- coreference-level features;
- features introduced for spoken language.

NP-level Features

1.	ante_gram_func	grammatical function of antecedent
2.	ante_npform	form of antecedent
3.	ante_agree	person, gender, number
4.	ante_case	grammatical case of antecedent
5.	ante_s_depth	the level of embedding in a sentence
6.	ana_gram_func	grammatical function of anaphor
7.	ana_npform	form of anaphor
8.	ana_agree	person, gender, number
9.	ana_case	grammatical case of anaphor
10.	ana_s_depth	the level of embedding in a sentence

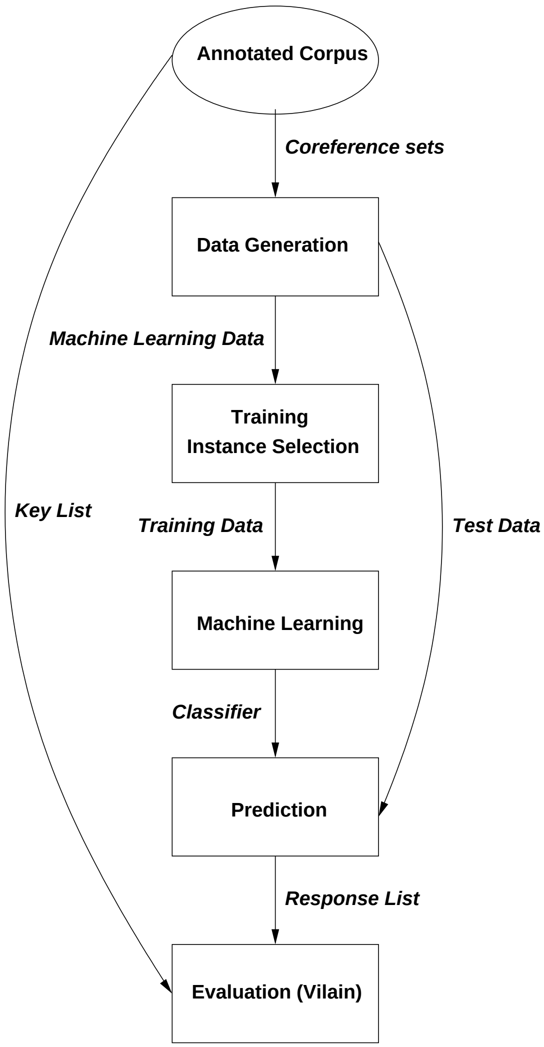
Coreference-level Features

11.	agree_comp	compatibility in agreement between anaphor and antecedent
12.	npform_comp	compatibilty in NP form between anaphor and antecedent
13.	wdist	distance between anaphor and antecedent in words
14.	mdist	distance between anaphor and antecedent in markables
15.	sdist	distance between anaphor and antecedent in sentences
16.	syn_par	anaphor and antecedent have the same grammatical function (yes, no)

Features for Dialogue

17.	ante_exp_type	type of antecedent (NP, S, VP)
18.	ana_np_pref	preference for NP arguments
19.	ana_vp_pref	preference for VP arguments
20.	ana_s_pref	preference for S arguments
21.	mdist_3mf3p	distance in NP-markables
22.	mdist_3n	distance in NP plus non-NP markables
23.	ante_tfidf	preference for *important* antecedents
24.	ante_ic	preference for *important* antecedents
25.	wdist_ic	distance is sum of IC of every word divided by number of words

Experimental Setup



Experimental Setup

- CART decision trees (R reimplementation: RPART)
  (R was chosen because it turned out to be a flexible environment without loss in speed compared to specialized software);
- all results reported obtained by 20-fold cross validation;
- baseline: NP-antecedents only; features as used for pronoun resolution in text;
- then *iterative procedure* applied for determining the best performing classifier and its features.

Iterative Procedure

(similar to *wrapper approach* for feature selection (Kohavi & John, 1997))

1. start with a model based on a set of predefined baseline features;

2. train models combining the baseline with all additional features seperately;

3. choose the best performing feature; add it to the model;

4. train models combining the enhanced model with each of the remaining features separately;

5. . . . repeat as long significant improvement can be observed.

Results: 3mf

	correct found	total found	total correct
baseline, features 1-16	120	150	1250
plus mdist_3mf3p	121	153	1250

	precision	recall	f-measure
baseline, features 1-16	80.00	9.60	17.14
plus mdist_3mf3p	79.08	9.68	17.25

Results: 3n

	correct found	total found	total correct
baseline, features 1-16	109	235	1250
plus none	97	232	1250
plus ante_exp_type	137	359	1250
plus wdist_ic	154	389	1250
plus ante_tfidf	158	391	1250

	precision	recall	f-measure
baseline, features 1-16	46.38	8.72	14.68
plus none	41.81	7.76	13.09
plus ante_exp_type	38.16	10.96	17.03
plus wdist_ic	39.59	12.32	18.79
plus ante_tfidf	40.41	12.64	19.26

Results: 3p

	correct found	total found	total correct
baseline, features 1-16	227	354	1250
plus wdist_ic	230	353	1250

	precision	recall	f-measure
baseline, features 1-16	64.12	18.16	28.30
plus wdist_ic	65.16	18.40	28.70

Results: Combined

	correct found	total found	total correct
baseline, features 1-16	456	739	1250
combined	509	897	1250

	precision	recall	f-measure
baseline, features 1-16	61.71	36.48	45.85
combined	56.74	40.72	47.42

Discussion

remaining problems:

- features which are supposed to prevent the resolution of pronouns without antecedents are not effective;
- identification of non-NP-antecedents difficult without semantic analysis;
- definite NPs, proper names should be included as well;
- binary classification is not optimal since the data contain to many negative cases (see Yang et al. (ACL 2003): competition learning approach);
- evaluation!

Conclusions: ML for Pronoun Resolution in Spoken Dialogue

- results comparable to Byron's (2002) who assumes semantic analysis and domain (in-)dependent knowledge; she also does not consider all pronouns;
- system ported successfully from anaphora resolution in written text to pronoun resolution in spoken dialogue;
- features derived from previous work do not work well, so, the theory did not keep it's promise.

Conclusions I

- in my work on anaphora resolution the development of theories precedes the implementation of systems;
- the development of theories includes – and builds upon – the annotation of corpora and their descriptive analysis;
- theories should always be tested against naturally occurring data, even if only hand-simulation is possible.

Conclusions II

However,

- I used different corpora for the development of the theory and for evaluating the system;
- theory and implemented system differ with respect to their expressiveness, coverage, evaluation methods;
- many features which were important for the theory do not show up in the final ML classifier;
- evaluation of the implemented system is much more rigorous.

Conclusions III

- go beyond simple anaphora resolution in written texts;
- distinguish between training and testing data;
- report results according to accepted evaluation methods;
- compare results to sensible baseline;
- try to publish at ACL (the reviewer’s comments are usually very good).

Further Information (Papers, Annotation Tool, . . .)

- `http://www.eml.org`
- `http://www.eml.org/nlp`
- `http://www.eml.org/english/homes/strube`
- email: `Michael.Strube@eml.villa-bosch.de`