University of Zurich[UZH]

Institute of Computational Linguistics

Incremental Pronoun Resolution in German with Markov Logic Networks

Don Tuggener
tuggener@cl.uzh.ch

7. Oktober 2014

Outline

1. Coreference resolution models: Mention-pair vs. Entity-mention
   - Implications for German pronoun resolution
2. Markov Logic Networks (MLNs) for German 3rd person pronoun resolution
3. Evaluation: compare rule-based vs. TiMBL vs. MLNs

Task definition

Coreference Resolution (CR) is the task of identifying and linking text surface expressions (*mentions*) that refer to the same underlying entity.

Example

US secretary of state, Hillary Clinton, trails Nancy Pelosi , speaker of the house of representatives, at number 36. Clinton, who days ago put a Congolese student in his place for asking what her husband Bill Clinton thought about a foreign policy issue, might have expected to be placed higher.

Task definition

Input: Text

Example

US secretary of state, Hillary Clinton, trails Nancy Pelosi , speaker of the house of representatives, at number 36. Clinton, who days ago put a Congolese student in his place for asking what her husband Bill Clinton thought about a foreign policy issue, might have expected to be placed higher.

Output: Coreference Partition
{ { US secretary of state, Hillary Clinton - Clinton - who - her}
{ a Congolese student - his } }

Standard Model: Mention-pair

Two step architecture

1. Identify local coreferring pairs of mentions {A-B}, {B-C}
   {Clinton - who}, {who - her}
2. Transitively merge pairs to chains (Best first) {A-B-C}
   {Clinton - who - her}

```
pairs = []
for NP in NPs:
 ante_cands = generate_candidates(NPs)
 ante = get_best(ante_cands)
 if ante:
  pairs.append([ante,NP])
coref_partition = merge_pairs(pairs)
```

Problems in the standard approach

- **Underspecification** of antecedent candidates **in local contexts**:
  Clinton → *she*? Hillary: ✓, Bill: ✗

- **Inconsistencies** during transitive closure (merging pairs)
  { Hillary Clinton, Clinton }
  { Clinton, he }
  ⇒ *{ Hillary Clinton, Clinton, he }*

Our solution: An incremental entity-mention model

One step architecture

```
buffer = []
coref_partition = []
for NP in NPs:
 ante_cands = generate_candidates(coref_partition)
 ante_cands += generate_candidates(buffer)
 ante = get_best(ante_cands)
 if ante:
  disambiguate(NP)
  #Hillary Clinton – Clinton –> Clinton = fem.
  coref_partition.extend(ante,NP)
 else:
  buffer.append(NP)
```

Solutions to problems in standard approach

- *Clinton* → *she*? Bill: ✗, Hillary: ✓
  ⇒ We (ideally) know whether it's Bill or Hillary based on previous decisions

- *{ Hillary Clinton, Clinton }*
  ~~*Clinton → he*~~
  ~~*{ Hillary Clinton, Clinton, he }*~~
  ⇒ *Clinton* → *he* not considered, we know *Clinton* is feminine.

This talk

- Improve pronoun resolution within this framework
- Compare rule-based vs. TiMBL vs. Markov Logic Networks (MLNs)
- Standard features as base, two new ones (animacy, NE types), combination of features in MLNs

Markov Logic Networks

Features

- 1st order predicate logic + stochastic inference
- Attractive framework: Most CR system combine rules and machine learning

Advantages over “common” ML architectures (kNN, SVM, MaxEnt...)

- No more fixed-length feature vectors (no dummy values, less noise)
- Weighting templates, 1 formula yields many weights
- Model feature dependencies / combinations

MLN architecture

Components of MLNs, adaption for CR

- global constraints: transitivity, exclusiveness
  $coref(A,B) \wedge coref(B,C) \rightarrow coref(A,C)$
  $coref(A,B) \wedge \neg coref(B,C) \rightarrow \neg coref(A,C)$
- local predicates: mention features (distance, gram. labels, ...)
  $in\_sent(A,1); has\_POS(A,PPER); has\_gf(A,SUBJ)$
- hidden predicates: coreference relations (what needs to be learned/inferred)
  $coref(A,B)$

Are combined in local soft constraints (i.e. weighted formulas)
$w(s2-s1) : in\_sent(A,s1) \wedge in\_sent(P,s2) \rightarrow coref(A,P)$

MLN complexity

- Common CR approaches with MLNs model **whole documents** as instances I
  - No. of hidden predicates per I equals sum of the pairwise permutation of mentions n pertained in each chain $c_{i...m}$
  - $|hidden\_predicates| \in I = \sum_{c_i}^{c_m} \frac{n_i!}{(n_i-2)!}$
  - Formulas for transitivity, exclusiveness, morphological agreement
- Our approach: model **individual pronoun instances** as I
  - No. of hidden predicates per I is simply a pronoun
  - $|hidden\_predicates| \in I = 1$
  - Transitivity, exclusiveness, morphological agreement $\rightarrow$ rule-based incremental entity-mention model

Control over weight learning

Learning the NE type of antecedent candidate A for pronoun P

- TiMBL feature:
  - A is NE: $PER, ORG, LOC, ...$
  - A is NN/PPER/PPOSAT: $*$ (dummy place holder)
- MLN formula:
  $w(ne\_type) : has\_pos(A, NE) \wedge has\_ne\_type(A, ne\_type) \rightarrow anaphoric(A, P)$

Control over weight learning

Learning the NE type of antecedent candidate A for pronoun P
... for each pronoun type (personal, possessive, relative)

- TiMBL feature:
  - A is NE: $PER, ORG, LOC, ...$
  - A is NN/PPER/PPOSAT: $*$ (dummy place holder)
    $\rightarrow$ 1 separate classifier for each pronoun type
- MLN formula: add POS tag to weighting condition
  $w(ne\_type, pos) : has\_pos(P, pos) \wedge has\_pos(A, NE) \wedge$
  $has\_ne\_type(A, ne\_type) \rightarrow anaphoric(A, P)$

Example: What we get from TiMBL (Gain Ratio)

PPER	
ne_type	0.0531

PPOSAT	
ne_type	0.0409

PRELS	
ne_type	0.0084

$\Rightarrow$ Feature has no impact on performance

Example: What we get from MLNs

PPER	PER	0.554021
	ORG	-0.123277
	OTH	-0.325205
	GPE	-0.750404
	LOC	-0.794775
PPOSAT	PER	0.751789
	ORG	0.197988
	OTH	0.016194
	GPE	-0.195872
	LOC	-0.307105
PRELS	PER	0.024996
	ORG	-0.250696
	OTH	-0.566349
	LOC	-0.788733
	GPE	-2.802126

$\Rightarrow$ Feature has positive impact

MLNs: Relations between features

Pronouns governed by discourse connectors tend to bind to antecedents in the same sentence.

Peter laughed because he ...

MLNs: connect sentence distance to presence of discourse connector

$w(s2-s1) : in\_sent(A,s1) \wedge in\_sent(P,s2) \wedge has\_connector(P) \rightarrow coref(A,P)$

$w(s2-s1) : in\_sent(A,s1) \wedge in\_sent(P,s2) \wedge \neg has\_connector(P) \rightarrow coref(A,P)$

$\Rightarrow$ cannot be easily expressed in TiMBL-like frameworks (separate classifiers for each combination)

Evaluation

Compare rule-based baseline (CorZu), TiMBL, MLNs

Data

- TübaD/Z v.9 (20% test, 20% dev, 60% train)

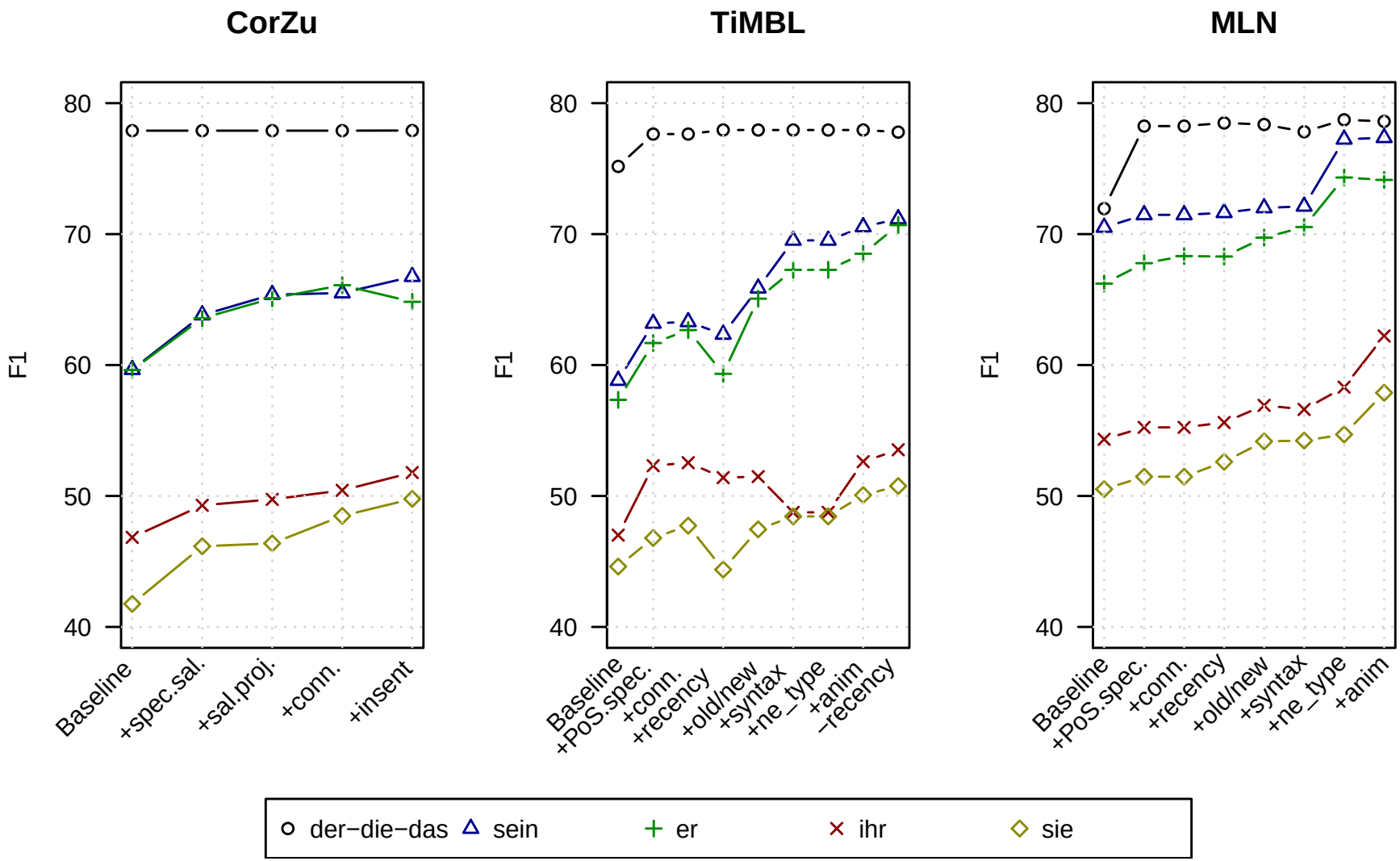
Metric

- CR / Pronoun resolution is a preprocessing step for higher-level applications
- Make metrics reflect what higher-level applications need
- E.g. require pronouns to (transitively) link to nominal antecedent

Evaluation ARCS F1 TübaD/Z test

	PPER		PPOSAT		PRELS
	sie	*er*	*sein*	*ihr*	*der|die|das*
CorZu	55.12	67.42	67.30	57.96	78.91
TiMBL	56.07	70.96	70.77	59.18	79.12
MLN	**62.17**	**75.41**	**74.44**	**65.10**	**79.46**

- MLNs perform best
- most recent baseline for relative pronouns good enough
- Large differences between pronoun types and lemmas
- masculine pronoun > fem./plural pronouns


CorZu
TiMBL
MLN
F1
80
70
60
50
40
Baseline
+spec.sal.
+sal.proj.
+conn.
+insent
+PoS.spec.
+recency
+old/new
+syntax
+ne_type
+anim
−recency
der−die−das
sein
er
ihr
sie

Evaluation F1 SemEval 2010 DE

	PPER		PPOSAT		PRELS
	sie	*er*	*sein*	*ihr*	*der/die/das*
BART	34.4	54.07	55.38	56.75	40.65
SUCRE	40.26	46.08	56.98	57.23	73.52
MLN	**53.33**	**64.94**	**73.66**	**65.07**	**80.39**

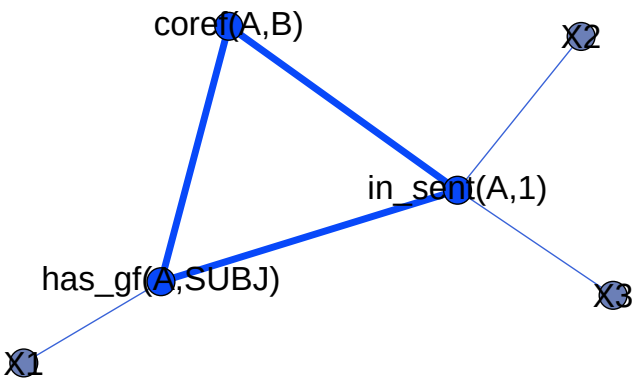
- Smaller TübaD/Z test set, fewer pronouns (≅200 per lemma, TübaD/Z >1000 per lemma)
- Other system were NOT specifically designed for pronoun resolution in German (multilingual coreference)

Conclusions

- Combination of incremental, discourse-aware text processing and MLNs for 3rd person pronoun resolution
- Outperforms all baselines by large margins
- MLNs can handle information which only applies to certain contexts (constraints) (c.f. NE types) and learn weights for each combination of feature value instantiations
- Novel features: NE types, Animacy
- New metric enabling detailed output analysis relevant for higher level applications
- Unveiled large performance differences for different pronoun types and lemmas (masc. > fem./pl.)

Learning/inferring with MLNs (light)

- Markov networks are graph models
- Binary (true/false) nodes represent predicates
- Formulas connect the predicate nodes and form cliques (containing local and the hidden predicate)



How often is a formula and its predicate instantiations true/false?

Evaluation Success Rate TübaD/Z test

Success rate: $\frac{|correctly\ resolved\ anaphors|}{|anaphors|}$

$\Rightarrow$ Any antecedent of the key is fine

	PPER		PPOSAT		PRELS
	sie	*er*	*sein*	*ihr*	*der|die|das*
CorZu	72.78	79.95	80.82	73.49	82.25
TiMBL	72.17	82.31	82.18	73.76	82.25
MLN	**77.74**	**85.64**	**86.14**	**80.08**	**82.82**

Evaluation Success Rate SemEval 2010 DE

	PPER		PPOSAT		PRELS
	sie	*er*	*sein*	*ihr*	*der/die/das*
BART	64.48	**79.33**	76.19	76.76	46.94
SUCRE	69.50	78.85	72.49	81.08	**81.11**
MLN	**76.83**	78.85	**82.01**	**85.41**	79.44

- Smaller test set, fewer pronouns (≅200 per lemma, TübaD/Z >1000 per lemma)
- Other system were NOT specifically designed for pronoun resolution in German (multilingual coreference)
- Large differences between pronoun types and lemmas
- masculine pronoun > fem./plural pronouns

Why is Pronoun Resolution in German so hard?

i.e. compared to English

DE	EN	Ambiguity
sie	*she/they*	number
ihr	*your/her/their*	number, person, (POS)
sein	*his/its*	gender

EN: *he, she* → person entities only
DE: *er, sie* → animate **AND** inanimate entities

⇒ More potential antecedent candidates
⇒ More room for errors
⇒ Entity-mention model handles ambiguity: *sein* → *er* ?

Evaluation Metrics for pronoun resolution

E.g. require pronouns to (transitively) link to nominal antecedent

Gold:	{ *Hillary Clinton*, *she*$_1$, *she*$_2$ }	Common	ARCS
Sys1:	{ *she*$_1$, *she*$_2$ }	R: 0.67	✗ R: 0
Sys2:	{ *Hillary Clinton*, *she*$_2$ }	R: 0.67	(✓) R: 0.5

Common metrics: Sys1==Sys2; 2 of 3 mentions found
ARCS: Sys1 no nominal antecedent; Sys2 found 1 of 2 anaphoric mentions

⇒ Harsh, but realistic metric for higher-level applications
⇒ Reports performance on any mention attribute, we look at lemmas

masc. > fem./pl.?

pos	*lemma*	*count*	*resolvable (%)*	*avg. antes*	*F1*
PPER	er	1146	93.46	4.72	75.41
PPER	sie	1177	86.40	5.01	62.17
PPOSAT	sein	812	94.82	7.93	74.44
PPOSAT	ihr	752	91.35	9.27	65.10
PRELS	der|die|das	1421	87.05	1.47	79.46

PPER		PPOSAT		PRELS	
most recent	0.2699	most recent	0.2967	most recent	0.5372
disc.stat.	0.1253	gf_parallel	0.0820	mable dist.	0.2353
sent. Dist.	0.0931	sent. Dist.	0.0792	disc.stat.	0.0501
pos_ante	0.0883	disc.stat.	0.0791	pos_ante	0.0361
pos_concat	0.0883	gf_ante	0.0747	pos_concat	0.0361
gf_ante	0.0858	gf_concat	0.0746	ne_class	0.0084
mable dist.	0.0786	pos_ante	0.0637	gf_ante	0.0053
gf_parallel	0.0722	pos_concat	0.0637	gf_concat	0.0030
ne_class	0.0531	mable dist.	0.0620	animacy	0.0001
gf_concat	0.0499	animacy	0.0493	gf_anaph	0.0001
animacy	0.0446	ne_class	0.0409	gf_parallel	0.0000
connector	0.0006	gf_anaph	0.0010	sent. Dist.	0.0000
gf_anaph	0.0001	connector	0.0000	connector	0.0000

Tabelle: Gain ratios of the features in the different TiMBL classifiers.

Anim.	Gender	Pos Tag	Weight
+	FEM	PPER	0.670906
+	MASC	PPER	0.455426
+	*	PPER	0.155619
+	*	PPOSAT	0.46031
-	NEUT	PRELS	0.475167
-	*	PRELS	0.170872
+	FEM	PRELS	0.098745
+	MASC	PRELS	0.084616
-	FEM	PPER	-0.670906
-	MASC	PPER	-0.455426
-	*	PPER	-0.155619
-	*	PPOSAT	-0.46031
+	NEUT	PRELS	-0.475167
+	*	PRELS	-0.170872
-	FEM	PRELS	-0.098745
-	MASC	PRELS	-0.084616

Tabelle: Weights for animacy of the antecedents based on gender.

Animacy

- NE types: Novel feature, pronouns tend to bind to PERSONs (Remember: In German, they can bind to other NE classes)
- Another novel feature: Animacy

Pos	Gender	Anim.	Weight
PPER	FEM	+	0.670906
PPER	MASC	+	0.455426
PPER	*	+	0.155619
PPOSAT	*	+	0.46031
PRELS	NEUT	-	0.475167
PRELS	*	-	0.170872
PRELS	FEM	+	0.098745
PRELS	MASC	+	0.084616

$\Rightarrow$ German pronouns tend to bind to animate entities