A Hybrid Entity-Mention Pronoun Resolution Model for German Using Markov Logic Networks

Don Tuggener
Unversity of Zurich
Institute of Computational Linguistics
tuggener@cl.uzh.ch

Manfred Klenner
Unversity of Zurich
Institute of Computational Linguistics
klenner@cl.uzh.ch

Abstract

This paper presents a hybrid pronoun resolution system for German. It uses a simple rule-driven entity-mention formalism to incrementally process discourse entities. Antecedent selection is performed based on Markov Logic Networks (MLNs). The hybrid architecture yields a cheap problem formulation in the MLNs w.r.t. inference complexity but pertains their expressiveness. We compare the system to a rule-driven baseline and an extension which uses a memory-based learner. We find that the MLN hybrid outperforms its competitors by large margins.

1 Introduction

Coreference resolution is an important tasks in many natural language processing pipelines. Several approaches have investigated the use of Markov Logic Networks (MLNs) for this task for the English language. Fewer approaches have explored MLNs for pronoun resolution, and, to our knowledge, none have explored the use of MLNs for German pronoun resolution.

We propose an architecture for the incorporation of MLNs in an entity-mention model for pronoun resolution in German. The hybrid architecture features two main benefits.

(i) The rule-driven, incremental entity-mention model provides a means to address the number of antecedent candidates, which is generally large in German due to morphological and semantic underspecification of certain key pronouns[1].

(ii) MLNs have attracted the attention of the coreference community, as global hard constraints can be used to enforce the transitivity and exclusiveness properties of coreference. Enforcing these properties poses problems in the classical mention-pair model (Soon et al., 2001, inter alia), where found pairs of coreferring NPs need to be merged to produce the coreference partition. The entity-mention model alleviates the need to express transitivity and exclusiveness in the MLNs, as the coreference partition is incrementally established during left-to-right processing and naturally adheres to these constraints. This allows us to model each pronoun occurrence as separate instance in the MLNs. Compared to other systems using MLNs, which model full documents, the hybrid architecture reduces the problem complexity for the MLN and, thereby, processing times.

We first review the incremental entity-mention model as implemented in the CorZu coreference system (Klenner and Tuggener, 2011). Next, we introduce the hybrid architecture which incorporates MLNs for antecedent selection. In the experiments section, we improve the CorZu system for pronoun resolution and establish a machine learning baseline based on TiMBL. Finally, we compare the three systems in the evaluation section[2].

[1]The pronoun *er (he)* can refer to both animate and inanimate entities, *sie (she/they)* is also underspecified in number; the possessive pronoun *sein* has ambiguous gender (masculine or neutral; *his/its*); the possessive pronoun *ihr* can be feminine, singular *(her)*, or plural *(their)*. Therefore, morphology cannot always be applied in a straight-forward way as a filter criterion for licensing antecedent candidates, which leads to large numbers of candidates.

2 Incremental discourse processing with an entity-mention model

To our knowledge, the CorZu system (Klenner and Tuggener, 2011) is the only ready-to-use system for coreference resolution for German[3]. The system implements a rule-driven entity-mention model, in which potential anaphors are compared to already established coreference sets and a buffer list which stores markables not yet in coreference sets. Algorithm 1 outlines the underlying discourse processing approach.

Algorithm 1 Incremental entity-mention model

```
1: for m ∈ Markables do
2:     for e ∈ CorefPartition do
3:         if e_{-1} < m ∧ compatible(e_{-1}, m) then
4:             Candidates ⊕ e_{-1}
5:         end if
6:     end for
7:     for np ∈ BufferList do
8:         if np < m ∧ compatible(np, m) then
9:             Candidates ⊕ np
10:        end if
11:    end for
12:    ante ← get_best(Candidates)
13:    if ∃ante then
14:        disambiguate(m)
15:        if ante ∈ CorefPartition then
16:            ante ⊕ m
17:        else
18:            CorefPartition ⊕ {ante ⊕ m}
19:        end if
20:    else
21:        BufferList ⊕ m
22:    end if
23: end for
```

For every markable[4] m, preceding markables are gathered from the coreference partition (lines 2-5; only the last mention of an established coreference chain is accessible, i.e. e_{-1}) and the buffer list (7-11) as antecedent candidates and appended ($\oplus$) to the candidate list. A selection strategy then determines the best candidate to be the antecedent (line 12). m is then disambiguated and absorbs number, gender, animacy, and named entity type of the antecedent (line 14). If the antecedent is a member of a coreference chain, m is appended to that chain (lines 15-16). Otherwise, a new coreference chain is created and appended to the coreference partition (lines 17-18). If no antecedent is determined, m is appended to the buffer list (line 21).

This architecture is attractive for pronoun resolution in German, because only one candidate (the most recent candidate e_{-1}) is accessible from discourse old entities (i.e. candidates already in coreference chains). When disambiguating a resolved mention, all semantic and morphological properties of the chain are projected onto e_{-1}[5]. Therefore, other members of the chain need not be considered as candidates when resolving successive markables, which potentially reduces the number of candidates from the length of a chain to one.

The main focus of the work presented here lies on improving the antecedent selection strategy (line 12) in algorithm 1 for pronoun resolution. The CorZu system uses a rule-based antecedent selection strategy based on a ranking of grammatical functions which determines the salience of the antecedent candidates[6]. The salience of each grammatical function gf is calculated by a simple ratio:

$$salience(gf) = \frac{|mentions\ bearing\ gf|}{|mentions|}$$

As the CorZu system was designed for general end-to-end coreference resolution, and not for pronouns in particular, we will experiment with rule-based extensions to this strategy. Before doing so, we will present the MLN based replacement of the antecedent selection strategy, which forms the main contribution of this work.

3 Markov Logic Networks for Reference Resolution

Markov Logic Networks (Richardson and Domingos, 2006) combine the strength of first order

[3] http://www.cl.uzh.ch/research/coreferenceresolution.html

[4] NPs potentially partaking in coreference relations.

[5] While we investigate the pronoun resolution component of CorZu in this work, the system still produces full coreference chains using string matching methods to link nominal mentions. We retain this mechanism to disambiguate potentially underspecified antecedent candidates.

[6] If two candidates have the same salience, the more recent one is selected.

predicate logic and stochastic inference. First order predicate formulas no longer need be binary, but can be assigned a weight based on statistical analysis of training data. MLNs are an interesting framework for coreference resolution, as most systems combine some notion of rule-based filtering and machine learning.

3.1 Related work

Song et al. (2012) propose a supervised model for coreference resolution using MLNs and compare it to a MaxEnt system under the same conditions. Their MLN system outperforms its MaxEnt variant and beats all other machine learning-based systems of the CoNLL 2011 shared task. Poon and Domingos (2008) investigate unsupervised coreference resolution with MLNs on the MUC-6 and ACE corpora and outperform the best results reported so far. Hou et al. (2013) apply MLNs to the problem of bridging anaphora. As Chan and Lam (2008) have shown, MLNs also provide a suitable framework for separately modeling pronoun resolution.

A strong motivation for using MLNs in coreference resolution in related work is that MLNs can be used to easily and efficiently address the problem of pair clustering. Transitivity and exclusiveness constraints can be expressed and enforced in simple first order predicate logic formulas.

3.2 Our approach

There are three types of formulas involved in modeling MLNs: local, global, and hidden ones. In coreference resolution, the local formulas are used to express soft constraints on the relation between pairs of mentions (e.g. sentence distance) which are assigned a weight during learning. Global hard constraints express the transitivity, symmetry, and exclusiveness properties of coreference and guide the pair clustering which generates the coreference partition. Finally, hidden predicates list the coreference relations between the mentions (i.e. the relations that need to be inferred during resolution).

A benefit of the entity-mention model is that clustering is not needed, as the coreference partition is established incrementally during left-to-right text processing, and the model naturally adheres to the transitivity and exclusiveness constraints of coreference. Therefore, we only need one global hard constraint, namely that a pronoun has exactly one antecedent.

As related work models whole documents in MLNs as instances, the number of hidden predicates per instance I is given by the number of mentions and the lengths of the chains they are in. This equals the sum of the pairwise permutation of mentions n pertained in each chain $c_{i...m}$, which amounts to

$$|hidden\_predicates| \in I = \sum_{c_i}^{c_m} \frac{n_i!}{(n_i-2)!}.$$

In contrast, because we do not need to express transitivity and exclusiveness in the MLN, we model each occurrence of a pronoun in a document as an instance and infer it separately, which gives us

$$|hidden\_predicates| \in I = 1.$$

Additionally, we reduce the MLN's workload by outsourcing the check for compatibility of antecedent candidates and a pronoun. Antecedent candidates are generated by the entity-mention model which uses hard filtering of candidates based on morphological agreement and distance[7]. This reduces the number of predicates and formulas needed in the MLN and, thereby, its complexity, which leads to fast processing times.

If clustering and, therefore, global constraints are not needed in our approach, the question why MLNs are still an interesting approach for this work arises. As e.g. Huang et al. (2009) noted, an important advantage of MLNs over other machine learning frameworks such as MaxEnt, kNN, Decision Trees, etc. is that weights are learned for instantiations of formulas, rather than for individual features. Similar to Conditional Random Fields, MLNs can express relations between features and weight them. Features are instantiated as predicates and be freely combined in formulas.

Furthermore, the weighting of formulas can be conditioned on any atom instantiated in the contained predicates. For example, conditioning the sentence distance between antecedents and pronouns on the pronoun type simply involves in-

[7]Relative pronouns can only have antecedents in the same sentence. Personal and possessive pronouns are allowed to have antecedents at most three sentences away. Unless a pronoun is underspecified in its morphological features, antecedent candidates must match in their morphology.

stantiating the PoS tag of the pronoun in a predicate and adding the tag to the weighting function. Such specification needs separate classifiers with specific training sets in other machine learning frameworks. Thus, MLNs provide an interesting framework, as different aspects of available information can be combined and weighted specifically.

4 Experiments

4.1 Data and evaluation metric

We use the TübaD/Z corpus (Hinrichs et al., 2005b) in its current version 9 for our experiments. The corpus contains 3444 newspaper articles annotated with coreference. We perform a 20%-20%-60% split on the data to obtain the test, development, and training sets[8]. Note that we use the gold preprocessing annotation throughout all our experiments to prevent preprocessing noise from influencing the comparison of the different approaches, but perform automated markable extraction. That is, we do not rely on the coreference annotation to identify which NPs should be considered as antecedent candidates.

As commonly used coreference metrics (MUC, BCUB, CEAF, BLANC) are not able report PoS-specific analysis of system outputs, they are not suited for pronoun resolution evaluation. Recently, Tuggener (2014) proposed the ARCS metrics, which are geared towards evaluation of coreference system outputs for higher level applications. These metrics provide PoS-based evaluation and can, therefore, be used for pronoun evaluation. Since the metrics can measure any annotated feature in corpus data, we report performance on the different pronoun types and their different lemmas[9].

The metrics use true positives (correctly resolved mentions), false negatives (unresolved mentions), and false positives (resolved markables that are not coreferential) to calculate Recall and Precision. The metrics also introduce a novel error class, called *wrong linkage*, which denotes coreferent mentions that have been resolved to wrong antecedents. Recall is calculated by $\frac{tp}{tp+wl+fn}$, and Precision by $\frac{tp}{tp+wl+fp}$. Recall thus extends over all mentions in the annotated corpus, and Precision calculation includes all coreference relations in the system output.

We choose the *ARCS inferred antecedent* metric which requires mentions to link to correct nominal antecedents within the coreference chain they are assigned to in order to be counted as true positives. The metric is strict in the sense that it does not reward simply linking pronouns to other pronouns. Only when pronouns (transitively) link to correct nominal antecedents they are regarded as true positives. We choose this metric, because we believe that pronoun resolution should at least infer correct local nominal antecedents in order to facilitate text understanding.

4.2 Extending the rule-based system

To establish a solid rule-based baseline, we add several constraints on the antecedent candidate generation mechanics in CorZu and report their impact on the development set in table 1.

PoS specific salience (+spec.sal.): The ranking of grammatical functions is performed uniformly for personal and possessive pronouns in CorZu. For relative pronouns, the most recent antecedent candidate is selected. We recalculate the salience of grammatical functions separately for personal and possessive pronouns to obtain pronoun type-specific salience rankings of the grammatical functions.

Grammatical function projection (+sal.proj.): The salience of an antecedent candidate is defined solely by the grammatical function it bears. From the discourse old entities, only the most recent mention is accessible for subsequent reference. Therefore, the grammatical function of the most recent mention of the entity determines its salience. We found that this is problematic when possessive pronouns are the most recent mentions, as they always bear the label *DET* (determiner), which is not as salient as e.g. *SUBJECT*. Therefore, if a possessive pronoun selects an antecedent within the same sentence (and is subsequently the only accessible

[8]For reproducibility, we report document ids: test set: text_0-text_689; dev set: text_690-text_1380; train set: text_1381-text_3444.

[9]We exclude the notorious (because potentially pleonastic) neutral pronoun *es (it)* from our experiments. We found that only around 10% of them are annotated as being anaphoric in the corpus. The baseline for not resolving *es* it therefore simply too high.

	Personal Pronouns						Possessive Pronouns						Relative Pronouns		
Lemma	sie			er			sein			ihr			der — die — das		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
baseline	**43.12**	**40.50**	**41.76**	**60.33**	**58.93**	**59.62**	**61.16**	**58.22**	**59.65**	**48.58**	**45.21**	**46.84**	**79.24**	**76.60**	**77.90**
+spec.sal.	47.50	44.90	46.16	64.44	62.72	63.57	65.47	62.25	63.82	51.06	47.63	49.29	79.24	76.60	77.90
+sal.proj.	47.71	45.13	46.38	66.00	64.23	65.10	67.08	63.77	65.38	51.54	48.07	49.74	79.24	76.60	77.90
+conn.	49.86	47.17	48.48	67.00	65.21	66.10	67.20	63.89	65.50	52.25	48.73	50.43	79.24	76.60	77.90
+insent	**51.29**	**48.36**	**49.78**	**65.72**	**63.97**	**64.83**	**68.43**	**65.14**	**66.75**	**53.78**	**49.95**	**51.79**	**79.23**	**76.63**	**77.91**

Table 1: Evaluation of extensions to the CorZu system on the development set.

mention of the entity), its grammatical function is overridden by that of the antecedent. Doing so, we prevent the salience of entities from being downgraded when they are referred to by a possessive pronoun in the same sentence.

Discourse connectors (+conn.): If a pronoun is preceded by a discourse connector, such as *because* or *although*, we only consider intra-sentential antecedent candidates. The intuition behind this constraint is that discourse relations such as *elaboration* or *contradiction* tend to have their arguments not too far apart in discourse. If a pronoun is an argument of such a relation, its antecedent should be nearby.

Intra-sentential candidates (+insent): A distance window of three sentences is often chosen to look for antecedents when resolving pronouns. However, pronouns tend to bind to intra-sentential antecedents quite frequently, disregarding the salience of the candidates. Therefore, we only keep candidates from within the same sentence, if available. Additionally, if there are pronouns among the intra-sentential candidates that are of the same PoS tag as the pronoun that is to be resolved, we discard all other candidates. Favoring the intra-sentential candidates is an attempt to complement the antecedent selection in CorZu, which is solely based on grammatical functions, with the similarly important factor of distance.

The results in table 1 show that all our extensions improve performance on personal and possessive pronouns. The relative pronouns do not seem to be affected, but their baseline performance is already quite strong. Calculating specific salience rankings of the grammatical functions for personal and possessive pronouns provides the highest single increase in performance. The other additions only marginally improve performance individually, but their cumulation leads to a solid upgrade of the CorZu system.

An interesting observation is the difference in performance regarding the gender of the personal and possessive pronouns. Performance on the masculine pronouns (*er, sein*) is much stronger. This may be caused by the fact that the feminine pronoun lemmas (*sie, ihr*) are ambiguous, i.e. they subsume the plural forms of the personal and possessive pronouns. These plural forms can have conjuncted NPs as antecedents, which are harder to handle.

4.3 TiMBL variant

To establish a machine learning-based baseline for the MLN system, we re-implement the TiMBL classifier approach by Klenner and Tuggener (2011). TiMBL is a kNN framework widely used in coreference and pronoun resolution (Hinrichs et al., 2005a; Hendrickx et al., 2007; Recasens and Hovy, 2009; Wunsch, 2010, inter alia). Klenner and Tuggener (2011) used individual classifiers for the different pronoun types. To stay close to their system, we implement three classifiers for each pronoun type, i.e. personal, possessive, and relative pronouns. The authors state that they used standard feature sets, but did not list them explicitly. In order to make available the same information to the TiMBL system as we will use in the MLN, we create the following feature vector for pairing an antecedent candidate i with a pronoun j:

baseline: Sentence and markable distance between i and j; grammatical function of i. **+syntax**: Grammatical function of j; whether the grammatical functions are parallel; concatenation of the grammatical functions of i and j; PoS tag of i. **+conn.**: Whether j is governed by a discourse connector. **+old/new**: Whether i is a new or old discourse entity (i.e. if the i stems from the

	Personal Pronouns						Possessive Pronouns						Relative Pronouns		
Lemma	sie			er			sein			ihr			der — die — das		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
baseline	**46.04**	**43.27**	**44.61**	**58.04**	**56.65**	**57.34**	**60.42**	**57.31**	**58.82**	**48.82**	**45.33**	**47.01**	**76.40**	**73.99**	**75.17**
+PoS.spec.	48.30	45.37	46.79	62.52	60.85	61.68	64.86	61.59	63.18	54.32	50.44	52.31	78.98	76.35	77.64
+conn.	49.27	46.28	47.73	63.53	61.83	62.67	64.98	61.71	63.30	54.56	50.66	52.54	78.98	76.35	77.64
+recency	*45.84*	*43.01*	*44.38*	*60.15*	*58.54*	*59.33*	*64.00*	*60.77*	*62.34*	*53.37*	*49.56*	*51.40*	79.29	76.65	77.95
+old/new	48.82	46.15	47.45	66.00	64.18	65.07	67.69	64.14	65.87	53.43	49.67	51.48	79.29	76.65	77.95
+syntax	50.00	46.97	48.44	68.19	66.37	67.27	71.15	67.96	69.52	50.65	46.98	48.75	79.29	76.65	77.95
+ne_type	50.00	46.97	48.44	68.19	66.37	67.27	71.15	67.96	69.52	50.65	46.98	48.75	79.29	76.65	77.95
+anim	51.67	48.56	50.07	69.38	67.65	68.50	72.38	68.82	70.55	54.67	50.71	52.62	79.29	76.65	77.95
-recency	**52.40**	**49.22**	**50.76**	**71.66**	**69.75**	**70.69**	**72.87**	**69.53**	**71.16**	**55.62**	**51.59**	**53.53**	**79.10**	**76.52**	**77.79**

Table 2: Evaluation of the TiMBL variant on the development set.

coreference partition or the buffer list). **+recency**: Whether i is the most recent candidate. **+anim.**: Animacy[10] of i. **+ne_type**: Named entity class of i.

Results of the TiMBL extension on the development set are shown in table 2. Note that we use the CorZu base system for processing, i.e. we remove the added rules from the previous experiment. For the **baseline**, we train a single classifier for all pronoun types. Next, we train separate classifiers for each pronoun type using the baseline features **(+PoS.spec.)**. We then incrementally add the additional features outlined above. To obtain the final TiMBL-based system, we remove the recency feature, as it impoverishes performance (−**recency**).

Evaluation shows that the TiMBL extension outperforms its rule-based counterpart especially for the masculine pronouns, and by a small difference in the female/plural pronouns. The biggest overall improvement stems from using separate, pronoun type-specific classifiers. Additionally, a relatively large performance increase can be observed for the masculine pronouns when adding the **+syntax** features, and the feminine/plural pronouns benefit from the **+anim** feature, especially. Note that the **+ne_type** feature does not have any affect on performance. We will return to this issue in section 4.5.

[10]We determine animacy of named entities by a list of first names gathered from the internet. If a named entity includes a name from this list, we label it as animate. For common nouns, we query GermaNet (Hamp and Feldweg, 1997) to assess whether the noun is a hyponym of the synset *Mensch (Human)*.

4.4 MLN hybrid

Next, we replace the antecedent selection step in algorithm 1 by MLNs. Table 3 shows the predicate logic formulas we experiment with. Markables in a document are enumerated from left to right following text direction. m denotes the numeric ID of a specific antecedent candidate for a specific numeric pronoun ID p. M denotes the set of available candidates for a given pronoun p. For learning, the most recent true antecedent among the candidates (i.e. the hidden predicate) is labeled based on the gold standard annotation.

We use *thebeast*[11] (Riedel, 2008) for MLN modeling. We set *thebeast* to use Integer Linear Programs for representing ground Markov networks and couple it with the *gurobi* solver[12] and learn for five epochs.

As in the TiMBL experiment, we remove the extensions to CorZu and use its vanilla instantiation as our base. We start with the baseline which uses only the formulas for sentence distance, markable distance, and grammatical function of the antecedent and incrementally append the formulas described in table 3. To enable PoS-specific weighting (**+PoS.spec.**), the predicate has_pos is added to each formula. For example, the formula for sentence distance is extended to:
$w(s2 - s1, pos) : insentence(m, s1) \wedge insentence(p, s2) \wedge has\_pos(p, pos) \rightarrow anaphoric(p, m)$
A weight is thus learned specifically for the different instantiations of the atoms in the weight function. In the sentence distance formula, the

[11]`https://code.google.com/p/thebeast/`
[12]`http://www.gurobi.com/`

Hidden predicate
Predicate to be inferred by the MLN: $anaphoric(p, m)$
Global hard constraint formula
The pronoun must have exactly one antecedent: $\|\forall m \in M : anaphoric(p, m)\| == 1$
Local soft constraint formulas
Distance-based formulas
-Sentence distance between m and p (**baseline**):
$w(s2 - s1) : insentence(m, s1) \wedge insentence(p, s2) \rightarrow anaphoric(p, m)$
-Markable distance if m and p are in the same sentence (**baseline**):
$w(p - m) : insentence(m, s) \wedge insentence(p, s) \rightarrow anaphoric(p, m)$
-Closest m to p (**+recency**):
$w : \|\forall m2 \in M : m2 > m\| == 0 \rightarrow anaphoric(p, m)$
-Closest m to p bearing "SUBJECT" as grammatical function (**+recency**):
$w : has\_gf(m, SUBJECT) \wedge \|\forall m2 \in M : has\_gf(m2, SUBJECT) \wedge m2 > m\| == 0 \rightarrow anaphoric(p, m)$
Syntax-based formulas
-Grammatical function of m (**baseline**):
$w(gf) : has\_gf(m, gf) \rightarrow anaphoric(p, m)$
-Parallelism of the grammatical functions of m and p (**+syntax**):
$w(gf) : has\_gf(m, gf) \wedge has\_gf(p, gf) \rightarrow anaphoric(p, m)$
-Transition of grammatical functions from m to p (**+syntax**):
$w(gf1, gf2) : has\_gf(m, gf1) \wedge has\_gf(p, gf2) \rightarrow anaphoric(p, m)$
Semantic formulas
-Animacy of m (**+anim.**):
$w(anim, gen, pos) : has\_animacy(m, anim) \wedge has\_gender(p, gen) \wedge has\_pos(p, pos) \rightarrow anaphoric(p, m)$
-Named entity type of m (**+ne_type**):
$w(ne\_type) : has\_pos(m, NE) \wedge has\_ne\_type(m, ne\_type) \rightarrow anaphoric(p, m)$
Discourse-based formulas
-Selecting m based on its discourse status (i.e. discourse-new vs. discourse-old) (**+old/new**):
$w(ds) : has\_discourse\_status(m, ds) \rightarrow anaphoric(p, m)$
-Sentence distance if p is preceded by a discourse connector (**+conn.**):
$w(s2 - s1) : insentence(m, s1) \wedge insentence(p, s2) \wedge has\_connector(p) \rightarrow anaphoric(p, m)$

Table 3: First order predicate logic formulas for MLN-based pronoun resolution in German

first value for the weight condition is the return value of a function over two atoms (the subtraction of numeric sentence IDs) and the second a PoS tag. Note that we apply this extension to all formulas in table 3.

For **+conn.**, the formula for weighting sentence distance in the presence of a discourse connective is added. **+recency** signifies the addition of the two formulas for weighting the most recent candidate and the most recent candidate bearing the grammatical label *SUBJECT*. **+old/new** adds the formula for selecting a discourse-old vs. a discourse-new candidate. **+syntax** signifies the addition of the formulas capturing the parallelism between m and p, and the transition of grammatical functions from m to p. Parallelism of grammatical functions has been used in pronoun resolution systems dating back to (Lappin and Leass, 1994). Capturing the transitions of grammatical functions from m to p is motivated by Centering theory (Grosz et al., 1995), which formulates typical transitions of grammatical functions of re-occurring entities in coherent texts. **+ne_type** weights the named entity type of m, if it is a named entity. **+anim** adds the formula for weighting the animacy of m specifically for each pronoun type and gender combination.

The results in table 4 show that the added formulas slowly but steadily increase pronoun resolution performance. A big improvement for the masculine pronouns stems from the addition of the NE type formula. For the feminine/plural pronouns, the animacy formula constitutes the single most significant improvement. Overall, the MLN hybrid outperforms the other systems by large margins. The MLN baseline using only three formulas already outperforms the CorZu extended system. Relative pronouns are the exception. Apart from learning PoS specific weights, they are not affected by the added formulas.

Lemma	Personal Pronouns						Possessive Pronouns						Relative Pronouns		
	sie			er			sein			ihr			der — die — das		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
baseline	**52.16**	**48.99**	**50.52**	**67.18**	**65.28**	**66.22**	**72.38**	**68.74**	**70.51**	**56.45**	**52.36**	**54.33**	**73.14**	**70.79**	**71.95**
+PoS.spec.	53.13	49.90	51.47	68.74	66.84	67.78	73.37	69.67	71.47	57.40	53.24	55.24	79.59	76.95	78.25
+conn.	53.13	49.90	51.47	69.29	67.38	68.32	73.37	69.67	71.47	57.40	53.24	55.24	79.59	76.95	78.25
+recency	54.32	51.01	52.61	69.20	67.41	68.29	73.49	69.87	71.63	57.75	53.63	55.61	79.84	77.18	78.49
+old/new	55.92	52.52	54.17	70.75	68.74	69.73	73.86	70.22	72.00	59.17	54.82	56.92	79.72	77.06	78.37
+syntax	55.99	52.58	54.23	71.57	69.54	70.54	73.98	70.34	72.12	58.82	54.56	56.61	79.16	76.53	77.82
+ne_type	56.48	53.04	54.70	75.41	73.27	74.32	79.28	75.29	77.24	60.59	56.20	58.31	80.09	77.42	78.73
+anim	**59.75**	**56.12**	**57.88**	**75.23**	**73.09**	**74.14**	**79.41**	**75.41**	**77.36**	**64.62**	**60.00**	**62.22**	**79.96**	**77.30**	**78.61**
TiMBL	52.40	49.22	50.76	71.66	69.75	70.69	72.87	69.53	71.16	55.62	51.59	53.53	79.10	76.52	77.79
CorZu	51.29	48.36	49.78	65.72	63.97	64.83	68.43	65.14	66.75	53.78	49.95	51.79	79.23	76.63	77.91

Table 4: Experiments with the MLN-extended system on the development set.

4.5 Comparison on the test set

Finally, we compare the systems on our test set. Table 5 reports the results. The system ranking does not change. However, we note that all systems achieve higher scores, especially for the feminine/plural pronouns.

A reason for the better performance of the MLN system compared to the TiMBL variant lies in the way they perform learning. While TiMBL calculates Gain Ratio for each of the 13 features in each of the three classifiers, amounting to 39 weights, *thebeast* learns a weight for each instantiation of the 11 formulas, which leads to 326 weights. That is, *thebeast* is able to absorb and apply the provided information in a more specific and detailed way.

Another benefit of *thebeast* manifests in the impact of adding NE types as a feature. In *thebeast*, we can require the formula to trigger only when the antecedent candidate is actually a named entity, indicated by the predicate $has\_pos(m, NE)$. The weight learning for this formula will only be triggered if this constraint is satisfied. In TiMBL, where fixed-length feature vectors are required, we need to insert a dummy value for the NE type feature if the antecedent candidate is not a NE. This dummy value will then be accounted for during feature weighting. Our evaluation on the development set showed that NE type information leads to a strong improvement in the MLN system, while it does not affect the TiMBL variant.

For error analysis, we checked the different error types that the ARCS metric measures. We found that all the systems have roughly the same number of false negatives and false positives. The false negative and false positive counts are much lower than the true positive and wrong linkage counts. For example, the MLN hybrid has the following counts for the *sie* pronoun: tp: 742, wl: 391, fn: 31, fp: 90. Therefore, it seems that it is the difference in the counts of true positives and wrong linkages that drives the difference in performance. However, we note that all our systems have much higher false positive than false negative counts, which indicates that the systems tend to resolve too many pronouns. A manual inspection of the system outputs showed that the false positives stem from cataphoric pronouns (which our systems treat as anaphors), generic uses of pronouns (which are anaphoric but not coreferent), and annotation errors (i.e. mostly missing annotations of pronouns).

5 Comparison to Related Work

Hinrichs et al. (2005a) experimented with German pronoun resolution on the TübaD/Z corpus. They first re-implemented the approach by Lappin and Leass (1994) for German and then explored TiMBL as a machine learning framework, using features based on distance and grammatical functions. The TiMBL system outperformed the rule-based system slightly, as in our experiments.

We have used similar features, but have explored two semantics-based ones, additionally. With the exception of Kouchnir (2004), who uses the semantic classes *human, physical*, or *abstract*, we are, to our knowledge, the first to use animacy and NE types as features in pronoun resolution for German. These features proved to significantly

	Personal Pronouns						Possessive Pronouns						Relative Pronouns		
Lemma	sie			er			sein			ihr			der — die — das		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
TübaD/Z test set															
MLN	**63.75**	**60.67**	**62.17**	**75.87**	**74.96**	**75.41**	**77.19**	**71.87**	**74.44**	**67.99**	**62.44**	**65.10**	**81.60**	**77.43**	**79.46**
TiMBL	57.23	54.96	56.07	71.24	70.69	70.96	73.09	68.60	70.77	61.62	56.93	59.18	81.25	77.10	79.12
CorZu	56.44	53.85	55.12	67.74	67.10	67.42	69.67	65.09	67.30	60.42	55.69	57.96	81.03	76.90	78.91
SemEval test set															
MLN	**52.04**	**54.69**	**53.33**	**64.79**	**65.09**	**64.94**	**72.73**	**74.61**	**73.66**	**64.55**	**65.59**	**65.07**	**79.39**	**81.43**	**80.39**
SUCRE	35.88	45.85	40.26	42.92	49.73	46.08	52.04	62.96	56.98	53.51	61.49	57.23	72.50	74.57	73.52
BART	33.83	35.00	34.40	53.30	54.85	54.07	54.82	55.96	55.38	54.79	58.86	56.75	40.60	40.71	40.65

Table 5: Comparison of systems on the test sets.

boost performance in our experiments.

Wunsch et al. (2009) explored instance sampling to reduce the large number of (negative) instances when resolving German pronouns. They used standard features and compared TiMBL to a decision tree and a maximum entropy learner. Instead of (under)sampling, we use the incremental entity-mention model to address the problem of the large number of (negative) instances.

In contrast to the approaches above, we aimed at detailed evaluation of pronoun resolution in a setting driven towards usability for higher-level applications. Therefore, we have used the ARCS metric which requires the closest nominal antecedent chosen by our systems to be correct. Our analysis showed that performance varies strongly between pronoun types and lemmas. We found that resolution of masculine pronouns is better than that of their female/plural counterparts.

As we used a more recent version of the TübaD/Z, we could not directly compare our results to previous work. However, the SemEval 2010 shared task on coreference resolution in multiple languages (Recasens et al., 2010) featured German as a language, with data drawn from the TübaD/Z[13]. We applied the ARCS scorer to the response files of the two best performing systems for German, namely SUCRE (Kobdani and Schütze, 2010) and BART (Broscheit et al., 2010), to measure their performance on pronoun resolution. We re-trained the MLN system on the shared task training data. Since we use GermaNet and gazetteers to obtain animacy information, our system falls in the category *open/gold*, like BART, while SUCRE participated in the *closed/gold* setting. The MLN system clearly outperforms the other two (cf. Table 5), although we have to consider that these systems were designed for multilingual coreference resolution and were not tuned for pronoun resolution in German.

[13]The TübaD/Z version used for SemEval is significantly smaller than the current version 9. In our test set based on version 9, there are 3 to 5 times more pronouns than in the SemEval test set. We choose the newer version because it therefore is a more solid foundation for our evaluation.

6 Conclusion

We have investigated the integration of MLNs into a state-of-the-art rule-based entity-mention model for German pronoun resolution. An advantage of the hybrid architecture over related work using MLNs lies in the reduction of the workload for the MLNs.

We have compared the MLN extension to a rule-based antecedent selection baseline and a TiMBL variant. The MLN system clearly outperformed its competitors in our experiments.

Additionally, we have found that there are large performance differences between different pronoun types and lemmas. Our evaluation showed that pronoun resolution still leaves room for substantial improvements when we require nominal antecedents to be produced. To our knowledge, we are the first to report detailed 3rd person pronoun resolution results on the TübaD/Z 9.

Acknowledgements

We would like to thank Angela Fahrni for her inspiration for investigating MLNs and *thebeast* and for her help in doing so. Additionally, we are grateful for the excellent and helpful comments by the anonymous reviewers.

References

Samuel Broscheit, Massimo Poesio, Simone Paolo Ponzetto, Kepa Joseba Rodriguez, Lorenza Romano, Olga Uryupina, Yannick Versley, and Roberto Zanoli. 2010. Bart: A multilingual anaphora resolution system. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 104–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ki Chan and Wai Lam. 2008. Pronoun resolution with Markov logic networks. In Hang Li, Ting Liu, Wei-Ying Ma, Tetsuya Sakai, Kam-Fai Wong, and Guodong Zhou, editors, *Information Retrieval Technology, 4th Asia Infomation Retrieval Symposium, AIRS 2008, Harbin, China, January 15-18, 2008, Revised Selected Papers*, volume 4993 of *Lecture Notes in Computer Science*, pages 153–164. Springer.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. 1995. Centering: a framework for modeling the local coherence of discourse. *Comput. Linguist.*, 21(2):203–225, June.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Iris Hendrickx, Veronique Hoste, and Walter Daelemans. 2007. Evaluating hybrid versus data-driven coreference resolution. In Antnio Branco, editor, *Anaphora: Analysis, Algorithms and Applications*, volume 4410 of *Lecture Notes in Computer Science*, pages 137–150. Springer Berlin Heidelberg.

Erhard Hinrichs, Katja Filippova, and Holger Wunsch. 2005a. What treebanks can do for you: Rule-based and machine-learning approaches to anaphora resolution in German. *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT05)*.

Erhard W. Hinrichs, Sandra Kübler, and Karin Naumann. 2005b. A unified representation for morphological, syntactic, semantic, and referential annotations. In *ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, Ann Arbor*.

Yufang Hou, Katja Markert, and Michael Strube. 2013. Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 907–917, Atlanta, Georgia, June. Association for Computational Linguistics.

Shujian Huang, Yabing Zhang, Junsheng Zhou, and Jiajun Chen, 2009. *Research in Computing Science: Advances in Computational Linguistics*, volume 41, chapter Coreference Resolution using Markov Logic Networks, pages 157–168. National Polytechnic Institute, Mexico.

Manfred Klenner and Don Tuggener. 2011. An incremental entity-mention model for coreference resolution with restrictive antecedent accessibility. In *Recent Advances in Natural Language Processing (RANLP 2011)*, pages 178–185.

Hamidreza Kobdani and Hinrich Schütze. 2010. Sucre: A modular system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 92–95, Stroudsburg, PA, USA. Association for Computational Linguistics.

Beata Kouchnir. 2004. A machine learning approach to German pronoun resolution. In *Proceedings of the ACL 2004 Workshop on Student Research*, ACLstudent '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shalom Lappin and Herbert J Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20:535–561.

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov logic. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 650–659, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marta Recasens and Eduard Hovy. 2009. A deeper look into features for coreference resolution. In *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium on Anaphora Processing and Applications*, DAARC '09, pages 29–42, Berlin, Heidelberg. Springer-Verlag.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew Richardson and Pedro Domingos. 2006. Markov logic networks. *Machine Learning*, 62(1-2):107–136, February.

Sebastian Riedel. 2008. Improving the accuracy and efficiency of map inference for Markov logic. In *Proceedings of the 24th Annual Conference on Uncertainty in AI (UAI '08)*, pages 468–475.

Yang Song, Jing Jiang, Wayne Xin Zhao, Sujian Li, and Houfeng Wang. 2012. Joint learning for coreference resolution with Markov logic. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing*

and Computational Natural Language Learning, EMNLP-CoNLL '12, pages 1245–1254, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wee M. Soon, Hwee T. Ng, and Daniel. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, December.

Don Tuggener. 2014. Coreference resolution evaluation for higher level applications. In *14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 231–235. Association for Computational Linguistics.

Holger Wunsch, Sandra Kübler, and Rachael Cantrell. 2009. Instance sampling methods for pronoun resolution. In *Recent Advances in Natural Language Processing (RANLP 2009)*, pages 478–483.

Holger Wunsch. 2010. *Rule-based and Memory-based Pronoun Resolution for German: A Comparison and Assessment of Data Sources*. Ph.D. thesis, Universität Tübingen.