A Cross-Lingual ILP Solution to Zero Anaphora Resolution

Ryu Iida
Tokyo Institute of Technology
2-12-1, Ôokayama, Meguro,
Tokyo 152-8552, Japan
ryu-i@cl.cs.titech.ac.jp

Massimo Poesio
Università di Trento,
Center for Mind / Brain Sciences
University of Essex,
Language and Computation Group
massimo.poesio@unitn.it

Abstract

We present an ILP-based model of zero anaphora detection and resolution that builds on the joint determination of anaphoricity and coreference model proposed by Denis and Baldridge (2007), but revises it and extends it into a three-way ILP problem also incorporating subject detection. We show that this new model outperforms several baselines and competing models, as well as a direct translation of the Denis / Baldridge model, for both Italian and Japanese zero anaphora. We incorporate our model in complete anaphoric resolvers for both Italian and Japanese, showing that our approach leads to improved performance also when not used in isolation, provided that separate classifiers are used for zeros and for explicitly realized anaphors.

1 Introduction

In so-called 'pro-drop' languages such as Japanese and many romance languages including Italian, phonetic realization is not required for anaphoric references in contexts in which in English non-contrastive pronouns are used: e.g., the subjects of Italian and Japanese translations of *buy* in (1b) and (1c) are not explicitly realized. We call these non-realized mandatory arguments **zero anaphors**.

(1) a. [EN] [John]$_i$ went to visit some friends. On the way, [he]$_i$ bought some wine.
b. [IT] [Giovanni]$_i$ andò a far visita a degli amici. Per via, ϕ_i comprò del vino.
c. [JA] [John]$_i$-wa yujin-o houmon-sita. Tochu-de ϕ_i wain-o ka-tta.

The felicitousness of zero anaphoric reference depends on the referred entity being sufficiently salient, hence this type of data–particularly in Japanese and Italian–played a key role in early work in coreference resolution, e.g., in the development of Centering (Kameyama, 1985; Walker et al., 1994; Di Eugenio, 1998). This research highlighted both commonalities and differences between the phenomenon in such languages. Zero anaphora resolution has remained a very active area of study for researchers working on Japanese, because of the prevalence of zeros in such languages[1] (Seki et al., 2002; Isozaki and Hirao, 2003; Iida et al., 2007a; Taira et al., 2008; Imamura et al., 2009; Sasano et al., 2009; Taira et al., 2010). But now the availability of corpora annotated to study anaphora, including zero anaphora, in languages such as Italian (e.g., Rodriguez et al. (2010)), and their use in competitions such as SEMEVAL 2010 Task 1 on Multilingual Coreference (Recasens et al., 2010), is leading to a renewed interest in zero anaphora resolution, particularly at the light of the mediocre results obtained on zero anaphors by most systems participating in SEMEVAL.

Resolving zero anaphora requires the simultaneous decision that one of the arguments of a verb is phonetically unrealized (and which argument exactly–in this paper, we will only be concerned with subject zeros as these are the only type to occur in Italian) and that a particular entity is its antecedent. It is therefore natural to view zero anaphora resolution as a joint inference

[1] As shown in Table 1, 64.3% of anaphors in the NAIST Text Corpus of Anaphora are zeros.

task, for which Integer Linear Programming (ILP)–introduced to NLP by Roth and Yih (2004) and successfully applied by Denis and Baldridge (2007) to the task of jointly inferring anaphoricity and determining the antecedent–would be appropriate.

In this work we developed, starting from the ILP system proposed by Denis and Baldridge, an ILP approach to zero anaphora detection and resolution that integrates (revised) versions of Denis and Baldridge's constraints with additional constraints between the values of three distinct classifiers, one of which is a novel one for subject prediction. We demonstrate that treating zero anaphora resolution as a three-way inference problem is successful for both Italian and Japanese. We integrate the zero anaphora resolver with a coreference resolver and demonstrate that the approach leads to improved results for both Italian and Japanese.

The rest of the paper is organized as follows. Section 2 briefly summarizes the approach proposed by Denis and Baldridge (2007). We next present our new ILP formulation in Section 3. In Section 4 we show the experimental results with zero anaphora only. In Section 5 we discuss experiments testing that adding our zero anaphora detector and resolver to a full coreference resolver would result in overall increase in performance. We conclude and discuss future work in Section 7.

2 Using ILP for joint anaphoricity and coreference determination

Integer Linear Programming (ILP) is a method for constraint-based inference aimed at finding the values for a set of variables that maximize a (linear) **objective function** while satisfying a number of constraints. Roth and Yih (2004) advocated ILP as a general solution for a number of NLP tasks that require combining multiple classifiers and which the traditional pipeline architecture is not appropriate, such as entity disambiguation and relation extraction.

Denis and Baldridge (2007) defined the following object function for the joint anaphoricity and coreference determination problem.

$$\min \sum_{\langle i,j\rangle \in P} c^{C}_{\langle i,j\rangle} \cdot x_{\langle i,j\rangle} + c^{-C}_{\langle i,j\rangle} \cdot (1 - x_{\langle i,j\rangle}) + \sum_{j \in M} c^{A}_{j} \cdot y_j + c^{-A}_{j} \cdot (1 - y_j) \quad (2)$$

subject to

$$x_{\langle i,j\rangle} \in \{0,1\} \qquad \forall \langle i,j\rangle \in P$$
$$y_j \in \{0,1\} \qquad \forall j \in M$$

M stands for the set of mentions in the document, and P the set of possible coreference links over these mentions. $x_{\langle i,j\rangle}$ is an indicator variable that is set to 1 if mentions i and j are coreferent, and 0 otherwise. y_j is an indicator variable that is set to 1 if mention j is anaphoric, and 0 otherwise. The costs $c^{C}_{\langle i,j\rangle} = -log(P(\mathrm{COREF}|i,j))$ are (logs of) probabilities produced by an antecedent identification classifier with $-log$, whereas $c^{A}_{j} = -log(P(\mathrm{ANAPH}|j))$, are the probabilities produced by an anaphoricity determination classifier with $-log$. In the Denis & Baldridge model, the search for a solution to antecedent identification and anaphoricity determination is guided by the following three constraints.

Resolve only anaphors: if a pair of mentions $\langle i,j\rangle$ is coreferent ($x_{\langle i,j\rangle} = 1$), then mention j must be anaphoric ($y_j = 1$).

$$x_{\langle i,j\rangle} \leq y_j \qquad \forall \langle i,j\rangle \in P \quad (3)$$

Resolve anaphors: if a mention is anaphoric ($y_j = 1$), it must be coreferent with at least one antecedent.

$$y_j \leq \sum_{i \in M_j} x_{\langle i,j\rangle} \qquad \forall j \in M \quad (4)$$

Do not resolve non-anaphors: if a mention is non-anaphoric ($y_j = 0$), it should have no antecedents.

$$y_j \geq \frac{1}{|M_j|} \sum_{i \in M_j} x_{\langle i,j\rangle} \qquad \forall j \in M \quad (5)$$

3 An ILP-based account of zero anaphora detection and resolution

In the corpora used in our experiments, zero anaphora is annotated using as markable the first verbal form (not necessarily the head) following the position where the argument would have been realized, as in the following example.

(6) [Pahor]$_i$ è nato a Trieste, allora porto principale dell'Impero Austro-Ungarico.
A sette anni [vide]$_i$ l'incendio del Narodni dom,

The proposal of Denis and Baldridge (2007) can be easily turned into a proposal for the task of detecting and resolving zero anaphora in this type of data by reinterpreting the indicator variables as follows:

- y_j is 1 if markable j (a verbal form) initiates a verbal complex whose subject is unrealized, 0 otherwise;
- $x_{\langle i,j \rangle}$ is 1 if the empty mention realizing the subject argument of markable j and markable i are mentions of the same entity, 0 otherwise.

There are however a number of ways in which this direct adaptation can be modified and extended. We discuss them in turn.

3.1 Best First

In the context of zero anaphora resolution, the 'Do not resolve non-anaphors' constraint (5) is too weak, as it allows the redundant choice of more than one candidate antecedent. We developed therefore the following alternative, that blocks selection of more than one antecedent.

Best First (BF):

$$y_j \geq \sum_{i \in M_j} x_{\langle i,j \rangle} \qquad \forall j \in M \qquad (7)$$

3.2 A subject detection model

The greatest difficulty in zero anaphora resolution in comparison to, say, pronoun resolution, is zero anaphora detection. Simply relying for this on the parser is not enough: most dependency parsers are not very accurate at identifying cases in which the verb does not have a subject on syntactic grounds only. Again, it seems reasonable to suppose this is because zero anaphora detection requires a combination of syntactic information and information about the current context. Within the ILP framework, this hypothesis can be implemented by turning the zero anaphora resolution optimization problem into one with *three* indicator variables, with the objective function in (8). The third variable, z_j, encodes the information provided by the parser: it is 1 with cost $c_j^S = -log(P(SUBJ|j))$ if the parser thinks that verb j has an explicit subject with probability $P(SUBJ|j)$, otherwise it is 0.

$$\begin{aligned} \min \sum_{\langle i,j \rangle \in P} & c^C_{\langle i,j \rangle} \cdot x_{\langle i,j \rangle} + c^{-C}_{\langle i,j \rangle} \cdot (1 - x_{\langle i,j \rangle}) \\ & + \sum_{j \in M} c^A_j \cdot y_j + c^{-A}_j \cdot (1 - y_j) \\ & + \sum_{j \in M} c^S_j \cdot z_j + c^{-S}_j \cdot (1 - z_j) \quad (8) \end{aligned}$$

subject to

$$\begin{aligned} x_{\langle i,j \rangle} &\in \{0,1\} & \forall \langle i,j \rangle \in P \\ y_j &\in \{0,1\} & \forall j \in M \\ z_j &\in \{0,1\} & \forall j \in M \end{aligned}$$

The crucial fact about the relation between z_j and y_j is that a verb has either a syntactically realized NP or a zero pronoun as a subject, but not both. This is encoded by the following constraint.

Resolve only non-subjects: if a predicate j syntactically depends on a subject ($z_j = 1$), then the predicate j should have no antecedents of its subject zero pronoun.

$$y_j + z_j \leq 1 \qquad \forall j \in M \qquad (9)$$

4 Experiment 1: zero anaphora resolution

In a first round of experiments, we evaluated the performance of the model proposed in Section 3 on zero anaphora only (i.e., not attempting to resolve other types of anaphoric expressions).

4.1 Data sets

We use the two data sets summarized in Table 1. The table shows that NP anaphora occurs more frequently than zero-anaphora in Italian, whereas in Japanese the frequency of anaphoric zero-anaphors[2] is almost double the frequency of the remaining anaphoric expressions.

Italian For Italian coreference, we used the annotated data set presented in Rodriguez et al. (2010) and developed for the Semeval 2010 task 'Coreference Resolution in Multiple Languages' (Recasens et al., 2010), where both zero-anaphora and NP

[2] In Japanese, like in Italian, zero anaphors are often used non-anaphorically, to refer to situationally introduced entities, as in *I went to John's office, but they told me that he had left.*

					#instances (anaphoric/total)		
language	type	#docs	#sentences	#words	zero-anaphors	others	all
Italian	train	97	3,294	98,304	1,093 / 1,160	6,747 / 27,187	7,840 / 28,347
	test	46	1,478	41,587	792 / 837	3,058 / 11,880	3,850 / 12,717
Japanese	train	1,753	24,263	651,986	18,526 / 29,544	10,206 / 161,124	28,732 / 190,668
	test	696	9,287	250,901	7,877 / 11,205	4,396 / 61,652	12,273 / 72,857

In the 6th column we use the term 'anaphoric' to indicate the number of zero anaphors that have an antecedent in the text, whereas the total figure is the sum of anaphoric and *exophoric* zero-anaphors - zeros with a vague / generic reference.

Table 1: Italian and Japanese Data Sets

coreference are annotated. This dataset consists of articles from Italian Wikipedia, tokenized, POS-tagged and morphologically analyzed using TextPro, a freely available Italian pipeline (Pianta et al., 2008). We parsed the corpus using the Italian version of the DESR dependency parser (Attardi et al., 2007).

In Italian, zero pronouns may only occur as omitted subjects of verbs. Therefore, in the task of zero-anaphora resolution all verbs appearing in a text are considered candidates for zero pronouns, and all gold mentions or system mentions preceding a candidate zero pronoun are considered as candidate antecedents. (In contrast, in the experiments on coreference resolution discussed in the following section, all mentions are considered as both candidate anaphors and candidate antecedents. To compare the results with gold mentions and with system detected mentions, we carried out an evaluation using the mentions automatically detected by the Italian version of the BART system (I-BART) (Poesio et al., 2010), which is freely downloadable.[3]

Japanese For Japanese coreference we used the NAIST Text Corpus (Iida et al., 2007b) version 1.4β, which contains the annotated data about NP coreference and zero-anaphoric relations. We also used the Kyoto University Text Corpus[4] that provides dependency relations information for the same articles as the NAIST Text Corpus. In addition, we also used a Japanese named entity tagger, *CaboCha*[5] for automatically tagging named entity labels. In the NAIST Text Corpus mention boundaries are not annotated, only the heads. Thus, we considered as pseudo-mentions all *bunsetsu* chunks (i.e. base phrases in Japanese) whose head part-of-speech was automatically tagged by the Japanese morphological analyser *Chasen*[6] as either 'noun' or 'unknown word' according to the NAIST-jdic dictionary.[7]

For evaluation, articles published from January 1st to January 11th and the editorials from January to August were used for training and articles dated January 14th to 17th and editorials dated October to December are used for testing as done by Taira et al. (2008) and Imamura et al. (2009). Furthermore, in the experiments we only considered *subject* zero pronouns for a fair comparison to Italian zero-anaphora.

4.2 Models

In these first experiments we compared the three ILP-based models discussed in Section 3: the direct reimplementation of the Denis and Baldridge proposal (i.e., using the same constrains), a version replacing Do-Not-Resolve-Not-Anaphors with Best-First, and a version with Subject Detection as well.

As discussed by Iida et al. (2007a) and Imamura et al. (2009), useful features in intra-sentential zero-anaphora are different from ones in inter-sentential zero-anaphora because in the former problem syntactic information between a zero pronoun and its candidate antecedent is essential, while the latter needs to capture the significance of saliency based on Centering Theory (Grosz et al., 1995). To directly reflect this difference, we created two antecedent identification models; one for intra-sentential zero-anaphora, induced using the training instances which a zero pronoun and its candidate antecedent appear in the same sentences, the other for

[3] http://www.bart-coref.org/

[4] http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/corpus.html

[5] http://chasen.org~taku/software/cabocha/

[6] http://chasen-legacy.sourceforge.jp/

[7] http://sourceforge.jp/projects/naist-jdic/

inter-sentential cases, induced from the remaining training instances.

To estimate the feature weights of each classifier, we used MEGAM[8], an implementation of the Maximum Entropy model, with default parameter settings. The ILP-based models were compared with the following baselines.

PAIRWISE: as in the work by Soon et al. (2001), antecedent identification and anaphoricity determination are simultaneously executed by a single classifier.

DS-CASCADE: the model first filters out non-anaphoric candidate anaphors using an anaphoricity determination model, then selects an antecedent from a set of candidate antecedents of anaphoric candidate anaphors using an antecedent identification model.

4.3 Features

The feature sets for antecedent identification and anaphoricity determination are briefly summarized in Table 2 and Table 3, respectively. The agreement features such as NUM_AGREE and GEN_AGREE are automatically derived using TextPro. Such agreement features are not available in Japanese because Japanese words do not contain such information.

4.4 Creating subject detection models

To create a subject detection model for Italian, we used the TUT corpus[9] (Bosco et al., 2010), which contains manually annotated dependency relations and their labels, consisting of 80,878 tokens in CoNLL format. We induced an maximum entropy classifier by using as items all arcs of dependency relations, each of which is used as a positive instance if its label is subject; otherwise it is used as a negative instance.

To train the Japanese subject detection model we used 1,753 articles contained both in the NAIST Text Corpus and the Kyoto University Text Corpus. By merging these two corpora, we can obtain the annotated data including which dependency arc is subject[10]. To create the training instances, any pair of a predicate and its dependent are extracted, each of which is judged as positive if its relation is subject; as negative otherwise.

As features for Italian, we used lemmas, PoS tag of a predicate and its dependents as well as their morphological information (i.e. gender and number) automatically computed by TextPro (Pianta et al., 2008). For Japanese, the head lemmas of predicate and dependent chunks as well as the functional words involved with these two chunks were used as features. One case specially treated is when a dependent is placed as an adnominal constituent of a predicate, as in this case relation estimation of dependency arcs is difficult. In such case we instead use the features shown in Table 2 for accurate estimation.

feature	description
SUBJ_PRE	1 if subject is included in the preceding words of *ZERO* in a sentence; otherwise 0.
TOPIC_PRE*	1 if topic case marker appears in the preceding words of *ZERO* in a sentence; otherwise 0.
NUM_PRE (GEN_PRE)	1 if a candidate which agrees with *ZERO* with regards to number (gender) is included in the set of *NP*; otherwise 0.
FIRST_SENT	1 if ZERO appears in the first sentence of a text; otherwise 0.
FIRST_WORD	1 if the predicate which has *ZERO* is the first word in a sentence; otherwise 0.
POS / LEMMA / DEP_LABEL	part-of-speech / dependency label / lemma of the predicate which has *ZERO*.
D_POS / D_LEMMA / D_DEP_LABEL	part-of-speech / dependency label / lemma of the dependents of the predicate which has *ZERO*.
PATH*	dependency labels (functional words) of words intervening between a *ZERO* and the sentence head

The features marked with '*' used only in Japanese.

Table 3: Features for anaphoricity determination

4.5 Results with zero anaphora only

In zero anaphora resolution, we need to find all predicates that have *anaphoric* unrealized subjects (i.e. zero pronouns which have an antecedent in a text), and then identify an antecedent for each such argument.

The Italian and Japanese test data sets contain 4,065 and 25,467 verbal predicates respectively. The performance of each model at zero-anaphora detection and resolution is shown in Table 4, using recall

[8] http://www.cs.utah.edu/~hal/megam/

[9] http://www.di.unito.it/~tutreeb/

[10] Note that Iida et al. (2007b) referred to this relation as 'nominative'.

feature	description
HEAD_LEMMA	characters of the head lemma in *NP*.
POS	part-of-speech of *NP*.
DEFINITE	1 if *NP* contains the article corresponding to DEFINITE 'the'; otherwise 0.
DEMONSTRATIVE	1 if *NP* contains the article corresponding to DEMONSTRATIVE such as 'that' and 'this'; otherwise 0.
POSSESSIVE	1 if *NP* contains the article corresponding to POSSESSIVE such as 'his' and 'their'; otherwise 0.
CASE_MARKER**	case marker followed by *NP*, such as '*wa* (topic)', '*ga* (subject)', '*o* (object)'.
DEP_LABEL*	dependency label of *NP*.
COOC_MI**	the score of well-formedness model estimated from a large number of triplets $\langle$ *NP*, Case, Predicate$\rangle$.
FIRST_SENT	1 if *NP* appears in the first sentence of a text; otherwise 0.
FIRST_MENTION	1 if *NP* first appears in the set of candidate antecedents; otherwise 0.
CL_RANK**	a rank of *NP* in forward looking-center list based on Centering Theory (Grosz et al., 1995)
CL_ORDER**	a order of *NP* in forward looking-center list based on Centering Theory (Grosz et al., 1995)
PATH	dependency labels (functional words) of words intervening between a *ZERO* and *NP*
NUM_(DIS)AGREE	1 if *NP* (dis)agrees with *ZERO* with regards to number; otherwise 0.
GEN_(DIS)AGREE	1 if *NP* (dis)agrees with *ZERO* with regards to gender; otherwise 0.
HEAD_MATCH	1 if *ANA* and *NP* have the same head lemma; otherwise 0.
REGEX_MATCH	1 if the string of NP subsumes the string of ANA; otherwise 0.
COMP_MATCH	1 if *ANA* and *NP* have the same string; otherwise 0.

NP, *ANA* and *ZERO* stand for a candidate antecedent, a candidate anaphor and a candidate zero pronoun respectively. The features marked with '*' are only used in Italian, while the features marked with '**' are only used in Japanese.

Table 2: Features used for antecedent identification

	Italian						Japanese		
	system mentions			gold mentions					
model	R	P	F	R	P	F	R	P	F
PAIRWISE	0.864	0.172	0.287	0.864	0.172	0.287	0.286	0.308	0.296
DS-CASCADE	0.396	0.684	0.502	0.404	0.697	0.511	0.345	0.194	0.248
ILP	0.905	0.034	0.065	0.929	0.028	0.055	0.379	0.238	0.293
ILP +BF	0.803	0.375	0.511	0.834	0.369	0.511	0.353	0.256	0.297
ILP +SUBJ	0.900	0.034	0.066	0.927	0.028	0.055	0.371	0.315	0.341
ILP +BF +SUBJ	0.777	0.398	0.526	0.815	0.398	**0.534**	0.345	0.348	**0.346**

Table 4: Results on zero pronouns

/ precision / F over link detection as a metric (model theoretic metrics do not apply for this task as only subsets of coreference chains are considered). As can be seen from Table 4, the ILP version with Do-Not-Resolve-Non-Anaphors performs no better than the baselines for either languages, but in both languages replacing that constraint with Best-First results in a performance above the baselines; adding Subject Detection results in further improvement for both languages. Notice also that the performance of the models on Italian is quite a bit higher than for Japanese although the dataset is much smaller, possibly meaning that the task is easier in Italian.

5 Experiment 2: coreference resolution for all anaphors

In a second series of experiments we evaluated the performance of our models together with a full coreference system resolving all anaphors, not just zeros.

5.1 Separating vs combining classifiers

Different types of nominal expressions display very different anaphoric behavior: e.g., pronoun resolution involves very different types of information from nominal expression resolution, depending more on syntactic information and on the local context and less on commonsense knowledge. But the most common approach to coreference resolu-

tion (Soon et al., 2001; Ng and Cardie, 2002, etc.) is to use a single classifier to identify antecedents of all anaphoric expressions, relying on the ability of the machine learning algorithm to learn these differences. These models, however, often fail to capture the differences in anaphoric behavior between different types of expressions–one of the reasons being that the amount of training instances is often too small to learn such differences.[11] Using different models would appear to be key in the case of zero-anaphora resolution, which differs even more from the rest of anaphora resolution, e.g., in being particularly sensitive to local salience, as amply discussed in the literature on Centering discussed earlier.

To test the hypothesis that using what we will call *separated models* for zero anaphora and everything else would work better than *combined models* induced from all the learning instances, we manually split the training instances in terms of these two anaphora types and then created two classifiers for antecedent identification: one for zero-anaphora, the other for NP-anaphora, separately induced from the corresponding training instances. Likewise, anaphoricity determination models were separately induced with regards to these two anaphora types.

5.2 Results with all anaphors

In Table 5 and Table 6 we show the (MUC scorer) results obtained by adding the zero anaphoric resolution models proposed in this paper to both a combined and a separated classifier. For the separated classifier, we use the ILP+BF model for explicitly realized NPs, and different ILP models for zeros.

The results show that the separated classifier works systematically better than a combined classifier. For both Italian and Japanese the ILP+BF+SUBJ model works clearly better than the baselines, whereas simply applying the original Denis and Baldridge model unchanged to this case we obtain worse results than the baselines. For Italian we could also compare our results with those obtained on the same dataset by one of the two systems that participated to the Italian section of SEMEVAL, I-BART. I-BART's results are clearly better than those with both baselines, but also clearly inferior to the results obtained with our models. In particular, the effect of introducing the separated model with ILP+BF+SUBJ is more significant when using the system detected mentions; it obtained performance more than 13 points better than I-BART when the model referred to the system detected mentions.

[11] E.g., the entire MUC-6 corpus contains a grand total of 3 reflexive pronouns.

	Japanese					
	combined			separated		
model	R	P	F	R	P	F
PAIRWISE	0.345	0.236	0.280	0.427	0.240	0.308
DS-CASCADE	0.207	0.592	0.307	0.291	0.488	0.365
ILP	0.381	0.330	0.353	0.490	0.304	0.375
ILP +BF	0.349	0.390	0.368	0.446	0.340	0.386
ILP +SUBJ	0.376	0.366	0.371	0.484	0.353	0.408
ILP +BF +SUBJ	0.344	0.450	0.390	0.441	0.415	**0.427**

Table 6: Results for overall coreference: Japanese (MUC score)

6 Related work

We are not aware of any previous machine learning model for zero anaphora in Italian, but there has been quite a lot of work on Japanese zero-anaphora (Iida et al., 2007a; Taira et al., 2008; Imamura et al., 2009; Taira et al., 2010; Sasano et al., 2009). In work such as Taira et al. (2008) and Imamura et al. (2009), zero-anaphora resolution is considered as a sub-task of predicate argument structure analysis, taking the NAIST text corpus as a target data set. Taira et al. (2008) and Taira et al. (2010) applied decision lists and transformation-based learning respectively in order to manually analyze which clues are important for each argument assignment. Imamura et al. (2009) also tackled to the same problem setting by applying a pairwise classifier for each argument. In their approach, a 'null' argument is explicitly added into the set of candidate argument to learn the situation where an argument of a predicate is 'exophoric'. They reported their model achieved better performance than the work by Taira et al. (2008).

Iida et al. (2007a) also used the NAIST text corpus. They adopted the BACT learning algorithm (Kudo and Matsumoto, 2004) to effectively learn subtrees useful for both antecedent identification and zero pronoun detection. Their model drastically outperformed a simple pairwise model, but it is still performed as a cascaded process. Incorporating

	Italian											
	system mentions						gold mentions					
	combined			separated			combined			separated		
model	R	P	F	R	P	F	R	P	F	R	P	F
PAIRWISE	0.508	0.208	0.295	0.472	0.241	0.319	0.582	0.261	0.361	0.566	0.314	0.404
DS-CASCADE	0.225	0.553	0.320	0.217	0.574	0.315	0.245	0.609	0.349	0.246	0.686	0.362
I-BART	0.324	0.294	0.308	–	–	–	0.532	0.441	0.482	–	–	–
ILP	0.539	0.321	0.403	0.535	0.316	0.397	0.614	0.369	0.461	0.607	0.384	0.470
ILP +BF	0.471	0.404	0.435	0.483	0.409	0.443	0.545	0.517	0.530	0.563	0.519	0.540
ILP +SUBJ	0.537	0.325	0.405	0.534	0.318	0.399	0.611	0.372	0.463	0.606	0.387	0.473
ILP +BF +SUBJ	0.464	0.410	0.435	0.478	0.418	**0.446**	0.538	0.527	0.533	0.559	0.536	**0.547**

R: Recall, P: Precision, F: f-score, BF: best first constraint, SUBJ: subject detection model.

Table 5: Results for overall coreference: Italian (MUC score)

their model into the ILP formulation proposed here looks like a promising further extension.

Sasano et al. (2009) obtained interesting experimental results about the relationship between zero-anaphora resolution and the scale of automatically acquired case frames. In their work, their case frames were acquired from a very large corpus consisting of 100 billion words. They also proposed a probabilistic model to Japanese zero-anaphora in which an argument assignment score is estimated based on the automatically acquired case frames. They concluded that case frames acquired from larger corpora lead to better f-score on zero-anaphora resolution.

In contrast to these approaches in Japanese, the participants to Semeval 2010 task 1 (especially the Italian coreference task) simply solved the problems using one coreference classifier, not distinguishing zero-anaphora from the other types of anaphora (Kobdani and Schütze, 2010; Poesio et al., 2010). On the other hand, our approach shows separating problems contributes to improving performance in Italian zero-anaphora. Although we used gold mentions in our evaluations, mention detection is also essential. As a next step, we also need to take into account ways of incorporating a mention detection model into the ILP formulation.

7 Conclusion

In this paper, we developed a new ILP-based model of zero anaphora detection and resolution that extends the coreference resolution model proposed by Denis and Baldridge (2007) by introducing modified constraints and a subject detection model. We evaluated this model both individually and as part of the overall coreference task for both Italian and Japanese zero anaphora, obtaining clear improvements in performance.

One avenue for future research is motivated by the observation that whereas introducing the subject detection model and the best-first constraint results in higher precision maintaining the recall compared to the baselines, that precision is still low. One of the major source of the errors is that zero pronouns are frequently used in Italian and Japanese in contexts in which in English as so-called *generic they* would be used: "*I walked into the hotel and (they) said ..*". In such case, the zero pronoun detection model is often incorrect. We are considering adding a generic they detection component.

We also intend to experiment with introducing more sophisticated antecedent identification models in the ILP framework. In this paper, we used a very basic pairwise classifier; however Yang et al. (2008) and Iida et al. (2003) showed that the relative comparison of two candidate antecedents leads to obtaining better accuracy than the pairwise model. However, these approaches do not output absolute probabilities, but relative significance between two candidates, and therefore cannot be directly integrated with the ILP-framework. We plan to examine ways of appropriately estimating an absolute score from a set of relative scores for further refinement.

Finally, we would like to test our model with English constructions which closely resemble zero anaphora. One example were studied in the Semeval 2010 'Linking Events and their Participants in Discourse' task, which provides data about *null instan-*

tiation, omitted arguments of predicates like "*We arrived ϕ^{goal} at 8pm.*". (Unfortunately the dataset available for SEMEVAL was very small.) Another interesting area of application of these techniques would be VP ellipsis.

Acknowledgments

Ryu Iida's stay in Trento was supported by the Excellent Young Researcher Overseas Visit Program of the Japan Society for the Promotion of Science (JSPS). Massimo Poesio was supported in part by the Provincia di Trento Grande Progetto LiveMemories, which also funded the creation of the Italian corpus used in this study. We also wish to thank Francesca Delogu, Kepa Rodriguez, Olga Uryupina and Yannick Versley for much help with the corpus and BART.

References

G. Attardi, F. Dell'Orletta, M. Simi, A. Chanev, and M. Ciaramita. 2007. Multilingual dependency parsing and domain adaptation using desr. In *Proc. of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, Prague.

C. Bosco, S. Montemagni, A. Mazzei, V. Lombardo, F. Dell'Orletta, A. Lenci, L. Lesmo, G. Attardi, M. Simi, A. Lavelli, J. Hall, J. Nilsson, and J. Nivre. 2010. Comparing the influence of different treebank annotations on dependency parsing. In *Proceedings of LREC*, pages 1794–1801.

P. Denis and J. Baldridge. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *Proc. of HLT/NAACL*, pages 236–243.

B. Di Eugenio. 1998. Centering in Italian. In M. A. Walker, A. K. Joshi, and E. F. Prince, editors, *Centering Theory in Discourse*, chapter 7, pages 115–138. Oxford.

B. J. Grosz, A. K. Joshi, and S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.

R. Iida, K. Inui, H. Takamura, and Y. Matsumoto. 2003. Incorporating contextual cues in trainable models for coreference resolution. In *Proceedings of the 10th EACL Workshop on The Computational Treatment of Anaphora*, pages 23–30.

R. Iida, K. Inui, and Y. Matsumoto. 2007a. Zero-anaphora resolution by learning rich syntactic pattern features. *ACM Transactions on Asian Language Information Processing* (*TALIP*), 6(4).

R. Iida, M. Komachi, K. Inui, and Y. Matsumoto. 2007b. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceeding of the ACL Workshop 'Linguistic Annotation Workshop'*, pages 132–139.

K. Imamura, K. Saito, and T. Izumi. 2009. Discriminative approach to predicate-argument structure analysis with zero-anaphora resolution. In *Proceedings of ACL-IJCNLP, Short Papers*, pages 85–88.

H. Isozaki and T. Hirao. 2003. Japanese zero pronoun resolution based on ranking rules and machine learning. In *Proceedings of EMNLP*, pages 184–191.

M. Kameyama. 1985. *Zero Anaphora: The case of Japanese*. Ph.D. thesis, Stanford University.

H. Kobdani and H. Schütze. 2010. Sucre: A modular system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 92–95.

T. Kudo and Y. Matsumoto. 2004. A boosting algorithm for classification of semi-structured text. In *Proceedings of EMNLP*, pages 301–308.

V. Ng and C. Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th ACL*, pages 104–111.

E. Pianta, C. Girardi, and R. Zanoli. 2008. The TextPro tool suite. In *In Proceedings of LREC*, pages 28–30.

M. Poesio, O. Uryupina, and Y. Versley. 2010. Creating a coreference resolution system for Italian. In *Proceedings of LREC*.

M. Recasens, L. Màrquez, E. Sapena, M. A. Martí, M. Taulé, V. Hoste, M. Poesio, and Y. Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8.

K-J. Rodriguez, F. Delogu, Y. Versley, E. Stemle, and M. Poesio. 2010. Anaphoric annotation of wikipedia and blogs in the live memories corpus. In *Proc. LREC*.

D. Roth and W.-T. Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proc. of CONLL*.

R. Sasano, D. Kawahara, and S. Kurohashi. 2009. The effect of corpus size on case frame acquisition for discourse analysis. In *Proceedings of HLT/NAACL*, pages 521–529.

K. Seki, A. Fujii, and T. Ishikawa. 2002. A probabilistic method for analyzing Japanese anaphora integrating zero pronoun detection and resolution. In *Proceedings of the 19th COLING*, pages 911–917.

W. M. Soon, H. T. Ng, and D. C. Y. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

H. Taira, S. Fujita, and M. Nagata. 2008. A Japanese predicate argument structure analysis using decision lists. In *Proceedings of EMNLP*, pages 523–532.

H. Taira, S. Fujita, and M. Nagata. 2010. Predicate argument structure analysis using transformation based learning. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 162–167.

M. A. Walker, M. Iida, and S. Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–232.

X. Yang, J. Su, and C. L. Tan. 2008. Twin-candidate model for learning-based anaphora resolution. *Computational Linguistics*, 34(3):327–356.