Inter-sentential Coreferences in Semantic Networks: An Evaluation of Manual Annotation

Václav Novák*, **Keith Hall**†

*Institute of Formal and Applied Linguistics
Charles University in Prague, Czech Republic
novak@ufal.mff.cuni.cz

† Center for Language and Speech Processing
Johns Hopkins University, USA
keith_hall@jhu.edu

Abstract

We present an evaluation of inter-sentential coreference annotation in the context of manually created semantic networks. The semantic networks are constructed independently be each annotator and require an entity mapping priori to evaluating the coreference. We introduce a model used for mapping the semantic entities as well as an algorithm used for our evaluation task. Finally, we report the raw statistics for inter-annotator agreement and describe the inherent difficulty in evaluating coreference in semantic networks.

1. Introduction

This paper presents an analysis of inter-sentential coreference relationships encoded in a manually annotated semantic network. The MultiNet semantic network formalism (Helbig, 2006) forms the basis of the annotation task used in this work. Under this formalism, each sentence in a discourse contributes information to the entire semantic network. These partial semantic entities are joined together by way of coreference links between unique *concepts* (i.e. the objects which define unique entities). Our particular annotation task is the extension of the deep-syntactic annotation currently available in the Prague Dependency Treebank (PDT) (Hajič et al., 2006). Thus adding constraints to the annotation which we describe in this paper.

Our annotators have manually constructed semantic networks for individual sentences while also maintaining coreference links between sentences; thereby providing complete semantic networks for an entire discourse. While the notion of coreference is related to the traditional notion, the goal here is to ensure that a coherent semantic network is constructed. This means that the *concepts* (entity objects) being annotated as coreferent must be interpreted as unique entities. In the present work, we explore a technique which maps the nodes of the semantic networks annotated by two different annotators. We use this technique to analyze the quality of the coreference annotations.

The remainder of the paper is organized as follows. In Section 2. we introduce the theoretical background of the MultiNet semantic network framework. Section 3. presents a description of the data we used for the annotation task, the model used to obtain a mapping between the labeling of multiple annotators, and the algorithm used to identify agreement. An evaluation of the coreference annotations is presented in Section 4. Finally, a discussion about the difficulties in the annotation process as well as the evaluation is presented in section Section 5.

2. MultiNet Semantic Networks

The representational means of Multilayered Extended Semantic Networks (MultiNet), which are described in (Helbig, 2006), provide a universal formalism for the treatment of semantic phenomena of natural language. To this end, the MultiNet is a parsimonious representation which provides a graphical interpretation of the semantic interactions realized throughout a discourse. This has an additional advantage of making the MultiNet networks easier to interpret without extensive knowledge of the formalism and therefore simplifies the training of annotators.

In Figure 1, we present an example MultiNet annotation for the following sentence from the Wall Street Journal: **Stephen Akerfeldt, currently vice president finance, will succeed Mr. McAlpine.**

As this network is for a single sentence, there are explicit inter-sentential coreferences. There are, however, intra-sentential coreferences links, those directed edges which are labeled **EQU** (meaning there is equality between the nodes C75 and C77 as well as C81 and C4). The networks is interpreted as follows: arcs indicate a semantic relationship between nodes (or sets of nodes). There are over 60 different categories of semantic relationships which describe the interactions between *concepts*; the concepts are encoded as nodes. Each concept is decorated further with detailed information that describes the type of concept. Note that an arc can be treated as a concept if in fact the semantic relationship is being treated in the discourse as a concept

In addition to the intra-sentential semantic relationships, a concept may be linked to a previously used concept in the discourse. For example, *Mr. McAlpine* had been mentioned previously in the discourse where this sentence appeared. Note that the text associated with node C4 does not actually appear in this sentence. C4 is a node from a previous sentence which has been presented here to note the coreference. The annotator has placed an **EQU** arc between nodes C4 and C81 to indicate there is a coreference relationship, and that C4 preceeded C81 in the discourse (depicted by the

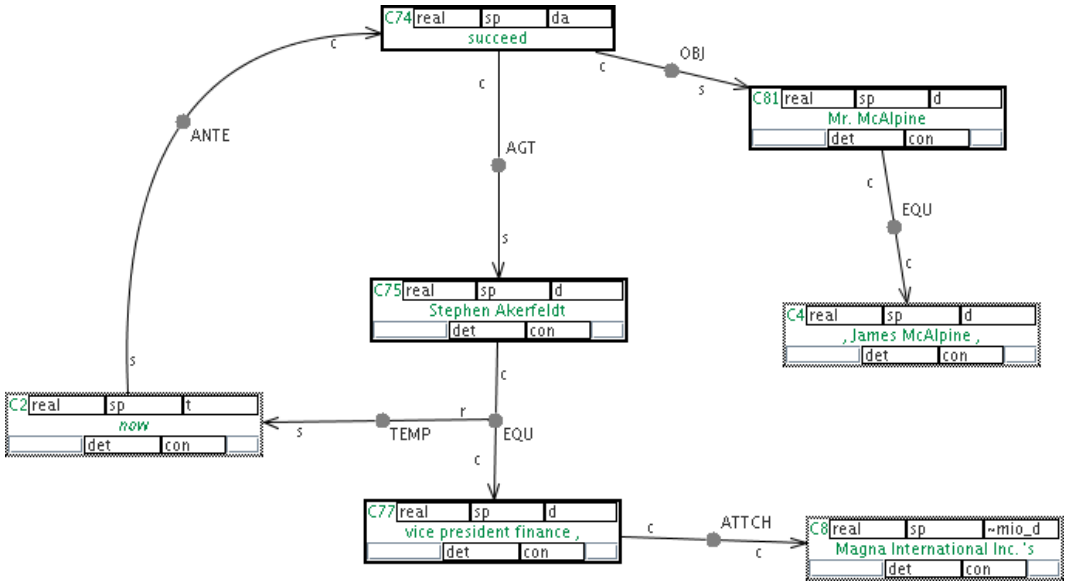

Figure 1: MultiNet annotation of the sentence "*Stephen Akerfeldt, currently vice president finance, will succeed Mr. McAlpine.*" Nodes C4 and C8 are re-used from previous sentences.

direction of the directed arc). These are the inter-sentential coreference links necessary to transform the set of localized semantic interactions into a complete semantic network that describes the relationships of concepts presented in a discourse.

A detailed annotation manual can be found at:

https://wiki.ufal.ms.mff.cuni.cz/projects:content-annotation

Manual annotations are performed using a graphical user interface, allowing the labeler to create nodes and arc and add labels to the nodes and arcs. Recall that our goal is to incorporate these annotations into the Prague Dependency Treebank. In order to facilitate this, the Tectogrammatical Representation (deep-syntax) trees of the PCEDT (PDT annotation of the Penn WSJ Treebank) are used to induce a default network. The nodes of the Tectogrammatical trees are directly mapped to MultiNet concepts, which may be further modified by the annotator. This imposes a loose constraint on the networks that are produced by the annotators: the networks will inherit structure from the Tectogrammatical trees.

In the following sections, we will examine the positive impact these annotation constraints have on the evaluation of networks annotated by multiple people.

3. Annotation: Data, Methodology, and Evaluation

The evaluation presented in this paper has been carried out on a subset of *The Wall Street Journal* articles from the Penn Treebank (Marcus et al., 1993), which have been annotated at multiple levels of analysis according to the PDT guidelines. The source data was publicly released as the Prague Czech-English Dependency Treebank (Cuřín et al., 2004), a corpus of parallel English and Czech annotations of the Penn Treebank. In this work, we have only explored the MultiNet annotation on the English component of this corpus.

The different levels of annotation are derived from the Functional Generative Description (FGD) of language (Sgall et al., 1986) which is the basis of the PDT annotation effort. This includes detailed morphological analysis, surface syntactic analysis, and deep-syntactic analysis; the latter is called the Tectogrammatical Representation (TR) as it was referred to in the original FGD work. Tectogrammatical trees are stripped of function words (synsemantic lexical items), leaving only the content bearing words as first-class nodes in the trees. Information derived from the function words is encoded in the labels and nodes of the TR dependency trees.

As mentioned above, the MultiNet annotation procedure begins with a TR tree, from which the annotator is given a default network that maintains links to the TR tree. The annotator then creates new nodes and arcs as necessary, and labels the nodes and arcs. Any concept, represented as a MultiNet node, which has previously been used in the discourse is available to the annotators via the annotation tool interface. When an annotator believe a concept in the current sentence has been previously mentioned, they can add this node to the current network (the node maintains it's identifier from the previous context). The annotator then creates an **EQU** link between the node for the previously used concept and the newly observed concept.

	Sentences	**Words**
Two-annotators	67	1793
Three-annotators	46	1236

Table 1: Annotated corpus. A subset of sentence from English side of the PCEDT (a PDT-style annotation of the WSJ treebank).

We trained three annotators to use the MultiNet graphical annotation tool (Novák, 2007). We reserved a set of sen-

tences from the corpus for training the annotators. These sentences are excluded for the inter-annotator analysis presented here. The complete evaluation for the three annotators contained 67 annotated sentences. Of these 67, only 46 sentences where annotated by all three annotators. The remaining sentences were annotated by only two of the annotators (see Table 1). All annotators are native English speakers and were trained on the held-out data.

The result of the annotation process is a set of manually annotated graphs. These graphs contain links to the tectogrammatical trees, which provide reference points for annotations from different annotators. Furthermore, the coreference link are maintained through the reuse of concept nodes when creating a graph. In order to evaluate agreement between two annotators, we first focus on the sentence-level graphs. We proceed as follows: first, we find an absolute mapping between the annotators for the MultiNet nodes. Then, we evaluate the coreference chains by identifying a canonical node that identifies the coreference chain.

3.1. Mapping Multiple Annotations

Mapping the MultiNet concepts as annotated by multiple annotators is necessary before further analysis can be done. The annotators are free to create new concept nodes if they deem it necessary to describes the semantic concepts within a sentence. One annotator may find it sufficient to use a node derived from the tectogrammatical tree, while another another may find the default node insufficient. In order to evaluate inter-annotator agreement for the network structure, we derived a technique to find an minimum-error-mapping for the nodes of a sentence. We employ the same technique in this work, where the goal is to evaluate coreference.

We used a relatively obvious technique for mapping the two graphs. First, all nodes that are derived from the tectogrammatical tree have an absolute identifier: the original tectogrammatical node. Therefore, the mapping between TR-derived nodes is fixed. The remaining concept nodes in the MultiNet tree are aligned in order to minimize the inter-annotator error for the graph annotation effort (independent of coreference annotations). It turns out that there are usually only a few new concept nodes created by the annotators for any one sentence, therefore, we can simply explore all mappings.

Formally, we start with a set of tectogrammatical trees containing a set of nodes N. The annotation is a tuple $G = (V, E, T, A)$, where V are the vertices, $E \subseteq V \times V \times P$ are the directed edges and their labels (e.g., agent of an action: AGT $\in P$), $T \subseteq V \times N$ is the mapping from vertices to the tectogrammatical nodes, and finally A are attributes of the nodes, which we ignore in this initial evaluation.[1] Analogously, $G' = (V', E', T', A')$ is another annotation of the same sentence and our goal is to quantify the differences between G and G'. This requires a mapping from V to V'. To find the optimal mapping we need a set Φ of admissible one to one mappings between vertices in the two annotations. A mapping is admissible if it connects vertices which are indicated by the annotators as representing the same tectogrammatical node:

$$\begin{aligned}\Phi \quad = \quad & \left\{ \phi \subseteq V \times V' \,\middle|\, \right. \\ & \underset{\substack{n\in N\\ v\in V\\ v'\in V'}}{\forall} \Big(\big((v,n)\in T \wedge (v',n)\in T'\big) \rightarrow (v,v')\in\phi \Big) \\ \wedge \quad & \underset{\substack{v\in V\\ v',w'\in V'}}{\forall} \Big(\big((v,v')\in\phi \wedge (v,w')\in\phi\big) \rightarrow (v'=w') \Big) \\ \wedge \quad & \left. \underset{\substack{v,w\in V\\ v'\in V'}}{\forall} \Big(\big((v,v')\in\phi \wedge (w,v')\in\phi\big) \rightarrow (v=w) \Big) \right\}\end{aligned} \tag{1}$$

In Equation 1, the first condition ensures that Φ is constrained by the mapping induced by the links to the tectogrammatical layer. The remaining two conditions guarantee that Φ is a one-to-one mapping.

Then we can define the optimal mapping ϕ^* as

$$\phi^* = \underset{\phi\in\Phi}{\operatorname{argmax}}(F(G, G', \phi)) \tag{2}$$

where F is similar to the F1-measure:

$$F_m(G, G', \phi) = \frac{2 \cdot m(\phi)}{|E| + |E'|} \tag{3}$$

where $m(\phi)$ is the number of edges that match given the mapping ϕ.

$$m(\phi) = |M_{dl}| + \frac{3}{4} \cdot |M_{wl}| + \frac{1}{2} \cdot |M_{dw}| + \frac{1}{4} \cdot |M_{ww}| \tag{4}$$

$|M_{dl}|$ is the number of edges where both direction and the label matches, $|M_{wl}|$ is the number of edges, where the direction is wrong but the label matches, $|M_{dw}|$ is the number of edges, where the direction is the same but the labels differ, and $|M_{ww}|$ is the number of edges, where the both direction and the label differ.

The coefficients in Equation 4 were chosen by hand in order to prefer mappings of the edges with more matching parameters and at the same time mappings where there are at least some structural correspondences. The relation type received more weight than the edge direction, because it is more informative. In the sequel, all results presented are obtained using the optimal mapping ϕ^* for each sentence.

Each concept (node) which occurs in more than one sentence is evaluated (these are the coreferent concepts which connect the sentence-level semantic networks). We choose a *canonical concept* for the one of the annotators by following the coreference chain to the earliest point at which the concept is mentioned in the discourse. This canonical concept is then identified in the second annotators graph. For every occurrence of the concept in each annotators graphs, we identify whether it is mapped under the previously describe sentence-level mapping. For any concept that is mapped, we identify whether it's canonical concept is also mapped. If so, this is a match under our metric. The complete algorithm is presented in Figure 2.

[1] We simplified the problem also by ignoring the mapping from edges to tectogrammatical nodes and the MultiNet edge attribute *knowledge type*.

Input: Alternative annotations, $G = (V, E, T, A)$ and $G' = (V', E', T', A')$
Output: List of coreference agreements and disagreements
foreach $v \in V$ *subject to* $|\{n \in N; (v, n) \in T\}| > 1$ **do**
    Find the first occurrence of the concept $n_0 \in N$, where $(v, n_0) \in T$;
    Find the $v'_0 \in V'$ such that $(v'_0, n_0) \in T'$;
    if *there is no such* v'_0 **then**
        `print("missingR0 for " + v);`
    else
        foreach $n \in N$ *where* $n \neq n_0 \wedge (v, n) \in T$ **do**
            Find the $v' \in V'$ such that $(v', n) \in T'$;
            if *there is no such* v' **then**
                `print("noMap for " + v + " at " + n);`
            else
                if $v' = v'_0$ **then**
                    `print("ok for " + v + " at " + n);`
                else
                    `print("mismatch of " + v' + " and " + v'_0);`

Figure 2: Comparing of two alternative coreference annotations. The asymptotic algorithmic complexity is $\mathcal{O}(|V| + |T|)$ because every inner loop iterates over different sets of n.

Annotator	Sentences	Unique concepts	(non-singleton)	Non-singletons per sent	(std. dev.)
SM	46	1248	(120)	2.6	(1.86)
CW	67	1713	(248)	3.7	(2.05)
CB	67	1800	(174)	2.6	(1.24)

Table 2: Annotation statistics for coreference evaluation. Non-singleton concepts are those with at least one coreference link.

4. Empirical Evaluation

We have run the coreference agreement evaluation on our annotated data, for which the relevant statistics are presented in Table 2. we report the raw results of the evaluation in Table 3. We have chosen not to report any further statistical evaluation due to 1) the limited amount of data available for analysis, and 2) the subtle dependency on the mapping procedure used as a basis for the analysis.
The results in are divided into four categories of coreferences:

mismatch One annotator uses a different canonical concept as the coreference target.

missingR0 The first mention of the concept in one annotator's graphs does not have a counterpart in the best matching network of the other annotator.

noMap The concept which is coreferring to a previous sentence in one annotators graph has no mapping to in the other annotators graph.

ok The coreferring concept used in the sentence by one annotator is mapped to a concept in the other annotator's graph and the canonical concepts from both annotators are mapped.

In Figure 3, we present a depiction of the agreement results that shows there is quite a bit of variance under the metric for **ok** agreements. Note that annotator SM appears to have less agreement in general.

5. Error Analysis

We have manually reviewed the disagreements as found using the above described metric. We have found that there are one of two explanations for many of the errors. One reason for disagreement appears to be an inadequate description of the coreference task in the annotation guidelines. The other error is related to the automatic mapping technique used in our evaluation.

The annotation guidelines do not indicate which previous concept to use when annotating coreference. An example of this was found for network annotations for sentences following this sentence from section F20 of the Penn WSJ Treebank: **The U.S. trade representative, Carla Hills, ...**
In the subsequent text, *Carla Hills* is used to identify the person. One annotator chose to use *Carla Hills* as the coreferent concept, but another chose to use *The U.S. trade representative*. The equality of a entities found in appositions of this sort can be resolved either by refining the annotation guideline or automatically preprocessing the data to identify appositive phrases.

The other source of error is related to the automatic mapping technique described in this paper. When one annotator's structural annotation is significantly different than the alternative annotation, the mapping algorithm will arbitrarily choose a mapping. This in turn misguides the coreference annotation algorithm.

Article	Status Pair	mismatch	missingR0	noMap	ok
F20	CB-CW	17	1	5	14
	SM-CB	15	0	24	11
	SM-CW	23	1	22	4
F21	CB-CW	3	0	6	8
	SM-CB	6	0	1	3
	SM-CW	5	0	2	3
F22	CB-CW	9	1	7	8
	SM-CB	7	1	0	3
	SM-CW	10	0	5	7
F26	CB-CW	6	0	5	7
F27	CB-CW	5	0	6	16

Table 3: Experimental results of pairwise coreference annotation agreement evaluation. The labels CB, CW, and SM, identify the individual annotators.

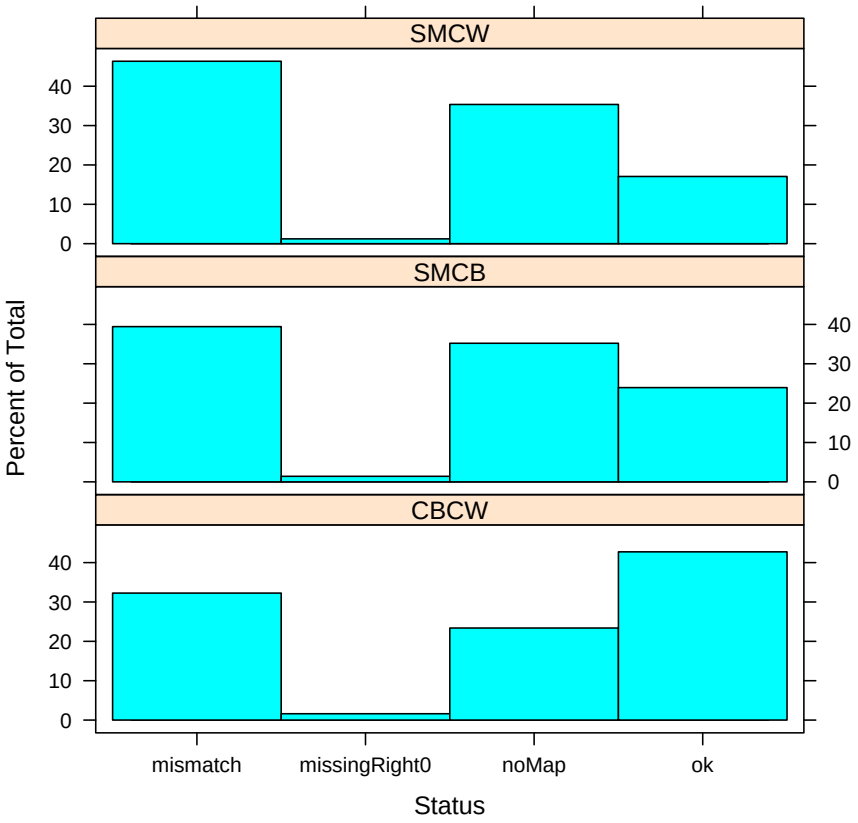

Figure 3: The agreement for all pairs of annotators, CBCW, SMCB, and SMCW. The data shows high variance w.r.t. the distribution of **ok** cases.

6. Conclusion and Future Work

We have presented a technique to evaluate coreference links in a semantic network annoation framework as well as the evaluation results on a small set of data annotated by three independent labalers. Evaluation under the current technique is inconclusive due to the complex nature of the annotation scheme and the integrated labeling of both structure an coreference structures.

Semantic network annotation is a relatively complex task which requires a high cognitive load even with the most parsimonious representations. Our preliminary results show that annotators are capable of producing similar annotations under the MultiNet representation. We hope that refinement in both the annotation guidelines and the evaluation technique will prove that MultiNet is an appropriate representation for high-agreement semantic network annotations.

We intend to use the proposed technique for subsequent coreference tasks in the role of consistency checking. We note that our technique is effective in determining that both the structural and coreference annotations agree.

Acknowledgment

This work was partially supported by Czech Academy of Science grants 1ET201120505 and 1ET101120503; by Czech Ministry of Education, Youth and Sports projects LC536 and MSM0021620838; and by the US National Science Foundation under grant OISE–0530118. The views expressed are not necessarily endorsed by the sponsors.

7. References

Jan Cuřín, Martin Čmejrek, Jiří Havelka, and Vladislav Kuboň. 2004. Building parallel bilingual syntactically annotated corpus. In *Proceedings of The First International Joint Conference on Natural Language Processing*, pages 141–146, Hainan Island, China.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, and Marie Mikulová. 2006. Prague Dependency Treebank 2.0. CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia, Pennsylvania.

Hermann Helbig. 2006. *Knowledge Representation and the Semantics of Natural Language*. Springer, Berlin, Germany.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–330.

Václav Novák. 2007. Cedit – semantic networks manual annotation tool. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 11–12, Rochester, New York, April. Association for Computational Linguistics.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel, Dordrecht, The Netherlands.