Aligning Predicate Argument Structures in Monolingual Comparable Texts: A New Corpus for a New Task

Michael Roth and **Anette Frank**
Department of Computational Linguistics
Heidelberg University
Germany
{mroth,frank}@cl.uni-heidelberg.de

Abstract

Discourse coherence is an important aspect of natural language that is still understudied in computational linguistics. Our aim is to learn factors that constitute coherent discourse from data, with a focus on how to realize predicate-argument structures (PAS) in a model that exceeds the sentence level. In particular, we aim to study the case of non-realized arguments as a coherence inducing factor. This task can be broken down into two subtasks. The first aligns predicates across comparable texts, admitting partial argument structure correspondence. The resulting alignments and their contexts can then be used for developing a coherence model for argument realization.

This paper introduces a large corpus of comparable monolingual texts as a prerequisite for approaching this task, including an evaluation set with manual predicate alignments. We illustrate the potential of this new resource for the empirical investigation of discourse coherence phenomena. Initial experiments on the task of predicting predicate alignments across text pairs show promising results. Our findings establish that manual and automatic predicate alignments across texts are feasible and that our data set holds potential for empirical research into a variety of discourse-related tasks.

1 Introduction

Research in the fields of discourse and pragmatics has led to a number of theories that try to explain and formalize the effect of discourse coherence inducing elements either locally or globally. For example, *Centering Theory* (Grosz et al., 1995) provides a framework to model local coherence by relating the choice of referring expressions to the salience of an entity at certain stages of a discourse. An example for a global coherence model would be *Rhetorical Structure Theory* (Mann and Thompson, 1988), which addresses overall text structure by means of *coherence relations* between the parts of a text.

In addition to such theories, computational approaches have been proposed to capture corresponding phenomena empirically. A prominent example is the entity-based model by Barzilay and Lapata (2008). In their approach, local coherence is modeled by the observation of sentence-to-sentence realization patterns of individual entities. The learned model reflects a key idea from Centering Theory, namely that adjacent sentences in a coherent discourse are likely to involve the same entities.

One shortcoming of Barzilay and Lapata's model (and extensions of it) is that it only investigates overt realization patterns in terms of grammatical functions. These functions reflect explicit realizations of predicate argument structures (PAS), but they do not capture the full range of salience factors. In particular, the model does not reflect the importance of discourse entities that fill core roles of the predicate, but that remain implicit in the predicate's local argument structure. We develop a specific set-up that allows us to further investigate the factors that govern such a null-instantiations of argument positions (cf. Fillmore et al. (2003)), as a special form of coherence inducing element in discourse. We henceforth refer to such cases as *non-realized arguments*.

Our main hypothesis is that context specific realization patterns for PAS can be automatically

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 218–227,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

learned from a semantically parsed corpus of comparable text pairs. This assumption builds on the success of previous research, where comparable and parallel texts have been exploited for a range of related learning tasks, e.g., unsupervised discourse segmentation (Barzilay and Lee, 2004) and bootstrapping semantic analyzers (Titov and Kozhevnikov, 2010).

For our purposes, we are interested in finding corresponding PAS across comparable texts that are known to talk about the same events, and hence involve the same set of underlying event participants. By aligning predicates in such texts, we can investigate the factors that determine discourse coherence in the realization patterns for the involved participants. As a first step towards this overall goal, we describe the construction of a resource that contains more than 160,000 document pairs that are known to talk about the same events and participants. Example (1), extracted from our corpus of aligned texts, illustrates this point: Both texts report on the same event, in particular the (<u>aligned</u>) event of locating victims in an avalanche. While (1.a) explicitly talks about the location of this event, the role remains implicit in the second sentence of (1.b), given that it can be recovered from the preceding sentence. In fact, realization of this argument would impede the fluency of discourse by being overly repetitive.

(1) a. ...The official said that [no bodies]$_{\text{Arg1}}$ had been <u>recovered</u> [from the avalanches]$_{\text{Arg2}}$ which occurred late Friday in the Central Asian country near the Afghan border some 300 kilometers (185 miles) southeast of the capital Dushanbe.

b. Three other victims were trapped *in an avalanche* in the village of Khichikh. [None of the victims bodies]$_{\text{Arg1}}$ have been <u>found</u> [ ]$_{\text{Argm-loc}}$.

Our aim is to identify comparable predications across pairs of texts, and to study the coherence factors that determine the realization patterns of argument structures (including roles that remain implicit) in discourse. This can be achieved by considering the full set of arguments that can be recovered from the aligned predications, including both core and non-core (i.e. adjunct) roles. However, in order to relate PAS across texts to one another, we first need to identify corresponding predicates.

In this paper, we construct a large data set to be used for the induction of a coherence model for argument structure realization and related tasks. We discuss the prospects of this data set for the study of coherence factors in PAS realization. Finally, we present first results on the initial task of *predicate alignment* across comparable monolingual texts.

The remainder of this paper is structured as follows: In Section 2, we discuss previous work in related tasks. Section 3 introduces the new task together with a description of how we prepared a suitable data set. Section 4 discusses the potential benefits of the created resource in more detail. Section 5 presents experiments on predicate alignment using this new data set and outlines first results. Finally, we conclude in Section 6 and discuss future work.

2 Related Work

Data sets comprising parallel texts have been released for various different tasks, including paraphrase extraction and statistical machine translation (SMT). While corpora for SMT are typically multilingual (e.g. Europarl, Koehn (2005)), there also exist monolingual parallel corpora that consist of multiple translations of one text into the same language (Barzilay and McKeown, 2001; Huang et al., 2002, inter alia). Each translation can provide alternative verbalizations of the same events but little variation can be observed in context, as the overall discourse remains the same. A higher degree of variation can be found in the Microsoft Research Paraphrase Corpus (e.g. MSRPC, Dolan and Brockett (2005)), which consists of paraphrases automatically extracted from different sources. In the MSRPC, however, original discourse contexts are not provided for each sentence. In contrast to truly parallel monolingual corpora, there also exist a range of comparable corpora that have been used for tasks such as (multi-document) summarization (McKeown and Radev, 1995, inter alia). Corpora for this task are collected manually and hence are rather small. Our work presents a method to automatically construct a large corpus of text pairs describing the same underlying events.

In this novel corpus, we identify common events across texts and investigate the argument structures that were realized in each context to establish a coherent discourse. Different aspects related to this setting have been studied in previous work. For example, Filippova and Strube (2007) and Cahill and Riester (2009) examine factors that determine constituent order and Belz et al. (2009) study the conditions for the use of different types of referring expressions. The specific set-up we examine allows us to further investigate the factors that govern the *non-realization* of an argument position, as a special form of coherence inducing element in discourse. As in the aforementioned work, we are specifically interested in the generation of coherent discourses (e.g. for summarization). Yet, our work also complements research in discourse analysis. A recent example for such work is the Semeval 2010 Task 10 (Ruppenhofer et al., 2010), which aims at linking events and their participants in discourse. The provided data sets for this task, however, are critically small (438 train and 525 test sentences). Eventually, the corpus we present in this paper could also be beneficial for data-driven approaches to role linking in discourse.

3 A Corpus for Aligning Predications across Comparable Texts

Our aim is to construct a corpus of comparable texts that can be assumed to be about the same events, but include variation in textual presentation. This requirement fits well with the news domain, for which we can trace varying textual sources for the same underlying events.

The English Gigaword Fifth Edition (Parker et al., 2011) corpus (henceforth just *Gigaword*) is one of the largest corpus collections for English. It comprises a total of 9.8 million newswire articles from seven distinct sources. For construction of our corpus we make use of all combinations of agency pairs in Gigaword.

3.1 Corpus Creation

In order to extract pairs of articles describing the same news event, we implemented the pairwise similarity method presented by Wubben et al. (2009). The method is based on measuring word overlap in news headlines, weighting each word by its TF*IDF score to give a higher impact to words occurring with lower frequency. As our focus is to provide a high-quality data set for predicate alignment and follow-up tasks, we impose an additional date constraint to favor precision over recall. We apply this constraint by requiring a pair of articles to be published within a two-day time frame in order to be considered as pairs of comparable news items.

Following this two-step procedure, we extracted a total of 167,728 document pairs, an overall collection of 50 million word tokens. We inspected about 100 randomly selected document pairs and found only two of them describing different events. This is in line with the results of Wubben et al. who reported a precision of 93% without explicitly imposing a date constraint. Overall, we found that most text pairs share a high degree of similarity and vary only in length (up to 7.564 words with a mean and median of 301 and 213 words, respectively) and detail. Closer examination of a development set of 10 document pairs (described below) revealed that we can indeed find multiple cases where roles are not locally filled in predicate argument structures. We show instances of this phenomenon, in which aligned PAS help to resolve implicit role references, in Section 4.

3.2 Gold Standard Annotation

We pre-processed all texts using MATE tools (Bohnet, 2010; Björkelund et al., 2010), a pipeline of natural language processing modules including a state-of-the-art semantic role labeler that computes Prop/NomBank annotations (Palmer et al., 2005; Meyers et al., 2008). The output was used to provide pre-labeled verbal and nominal predicates for annotation. We asked two students[1] to tag alignments of corresponding predicates in 70 text pairs derived from the created corpus. All document pairs were randomly chosen from the AFP and APW sections of Gigaword with the constraint that each text consists of 100 to 300 words[2]. We chose this constraint as longer text pairs contain a high number of unrelated predicates, making this task difficult to manage for the annotators.

Sure and possible links. Following standard practice in word alignment tasks (cf. Cohn et al. (2008))

[1] Both annotators are students in Computational Linguistics, one undergraduate (A) and one postgraduate (B) student.

[2] This constraint is satisfied by 75.3% of the documents.

the annotators were instructed to distinguish between *sure* (S) and *possible* (P) alignments, depending on how certainly, in their opinion, two predicates (including their arguments) describe the same event. The following examples show cases of predicate pairings marked as sure (S link) (2) and as possible (P link) alignments (3):

(2) a. The regulator ruled on September 27 that Nasdaq too was qualified to bid for OMX [...][3]

b. The authority [...] had already approved a similar application by Nasdaq.[4]

(3) a. Myanmar's military government said earlier this year it has released some 220 political prisoners [...][5]

b. The government has been regularly releasing members of Suu Kyi's National League for Democracy party [...][6]

Replaceability. As a guideline for deciding whether two predicates are to be aligned, the annotators were given the following two criteria: 1) whether the predicates are replaceable in a given context and 2) whether they share (potentially implicit) arguments.

Missing context. In case one text does not provide enough context to decide whether two predicates in the paired documents refer to the same event, an alignment should not be marked as sure.

Similar predicates. Annotators were told explicitly that sure links can be used even if two predicates are semantically different but have the same meaning in context. Example (4) illustrates such a case:

(4) a. The volcano roared back to life two weeks ago.

b. It began erupting last month.

1-to-1 vs. n-to-m. We asked the annotators to find as many 1-to-1 correspondences as possible and to prefer 1-to-1 matches over n-to-m alignments. In case of multiple mentions (cf. Example (5)) of the same event, we further asked the annotators to provide only one S link per predicate and mark remaining cases as P links. If possible, the S link should be used for the pairing of PAS with the highest information overlap (e.g. "perform$_{a3}$"–"perform$_{b2}$" in (5)). If there is no difference in information overlap, the predicate pair that occurs first in both texts should be marked as a sure alignment (e.g. "sing$_{a1}$"–"perform$_{b1}$" in (5)). The intuition behind this guideline is that the first mention introduces the actual event while later mentions just (co-)refer or add further information.

(5) a. Susan Boyle said she will sing$_{a1}$ in front of Britain's Prince Charles (...) "It's going to be a privilege to be performing$_{a2}$ before His Royal Highness," the singer said (...) British copyright laws will allow her to perform$_{a3}$ the hit in front of the prince and his wife.[7]

b. British singing sensation Susan Boyle is going to perform$_{b1}$ for Prince Charles (...) The show star will perform$_{b2}$ her version of Perfect Day for Charles and his wife Camilla.[8]

3.3 Development and Evaluation Data Sets

In total, the annotators (A/B) aligned 487/451 sure and 221/180 possible alignments with a Kappa score (Cohen, 1960) of 0.86. Following Brockett (2007), we computed agreement on labeled annotations, including unaligned predicate pairs as an additional *null* category. For the construction of a gold standard, we merged the alignments from both annotators by taking the union of all possible alignments and the intersection of all sure alignments. Cases which involved a sure alignment on which the annotators disagreed were resolved in a group discussion with the first author. We split the final corpus into a development set of 10 document pairs and a test set of 60 document pairs.

Table 1 summarizes information about the resulting annotations in the development and test sets, respectively. It gives information about the paired texts (PT): number of predicates marked in preprocessing (nouns and verbs), the set of manual predicate alignments (PA): sure and possible, as well as information about whether they were annotated for predicates of the same PoS (N,V) or lemma.

Finally, as a rough indicator for diverging argument structures captured in the annotated align-

[3] Source document ID: AFP_ENG_20071112.0235

[4] Source document ID: APW_ENG_20071112.0645

[5] Source document ID: AFP_ENG_20020301.0041

[6] Source document ID: APW_ENG_20020301.0132

[7] Source document ID: AFP_ENG_20101102.0028

[8] Source document ID: APW_ENG_20101102.0923

	Dev Set	**Test Set**
nb. of PT	10	60
nb. marked predicates	395	3,453
nb. marked nouns	168	1,531
nb. marked verbs	227	1,922
sure PA/PT: avg. (total)	3.9 (35)	7.4 (446)
poss. PA/PT: avg. (total)	4.8 (43)	6.0 (361)
same PoS in PA (N/V)	88.5% (24/42)	82.4% (242/423)
same lemma in PA	53.8% (42)	47.5% (383)
unequal nb. args in PA	30.8% (24)	39.7% (320)

Table 1: Information on Paired Texts (PT) and manual Predicate Alignments (PA) in development and test set

ments, we analyzed the number of PAs that involve a different number of arguments.

4 Potential of Aggregation

In this section, we analyze the predicate alignments in our manually annotated data set, to illustrate the potential of aggregating corresponding PAS across comparable texts.

We are particularly interested in cases of non-realization of arguments, and thus take a closer look at alignments involving roles that are not filled in their local PAS. We extract a subset of such cases by extracting pairs of aligned predicates that contain a different number of realized arguments. We deliberately focus on the more restricted core roles in this exposition, but will consider the full range of roles for developing a comprehensive coherence model for argument structure realization.[9] Our selection of alignment examples is drawn from the development set.

The following excerpts are from a pair of comparable texts describing a news report on Chadian refugees crossing into Nigeria:

(6) a. The Chadians said [they]$_{\text{Arg0}}$ had fled [ ]$_{\text{Arg1}}$ in fear of their lives.[10]

b. The United Nations says [some 20,000 refugees]$_{\text{Arg0}}$ have fled [into Cameroon]$_{\text{Arg1}}$.[11]

In both examples, the Arg0 role of the predicate fled is filled, but Arg1 has not been realized in (6.a). Note that the sentence is still part of a coherent discourse as fillers for the omitted role can be inferred from the preceding discourse context. Aggregating the aligned PAS presents an effective means to identify such appropriate fillers.

Example (7) presents another text pair, reporting on elections in Iraq, in which role realizations differ for the same hold event.

(7) a. He said (...) [elections]$_{\text{Arg1}}$ will be held [ ]$_{\text{Arg0}}$ to form a government.[12]

b. The president (...) said Wednesday [his country]$_{\text{Arg0}}$ will definitely hold [elections]$_{\text{Arg1}}$ in 2004.[13]

Here, the changes in argument realization go along with a diathesis alternation, while the pair in (6) exemplifies a case of lexical licensing for omission of a role.[14]

Example (8.b) illustrates a case in which the Arg1 of a decline event is involved in a preceding predication (*rise*) and thus has already been overtly realized. The constructional properties of the subsequent predicates *decline* as a participle and noun, respectively, are more adverse to overt realization of the Arg1 role. Suppression of Arg1 in such cases yields a much more coherent discourse as compared to their realization. This is brought out by the constructed examples in (a'/b'), which are both highly repetitive.

(8) a. The closely watched [index]$_{\text{Arg1}}$ rose to 93.7 ... after declining for ... months.[15]

a'. ? ... after the index declining for ... months.

b. Consumer confidence rose ... following three months of dramatic decline [ ]$_{\text{Arg1}}$.[16]

b'. ? ... following three months of dramatic decline [of consumer confidence]$_{\text{Arg1}}$.

As showcased by the previous examples, the decision on whether to realize a role filler in a local PAS can be rather complex. Obviously, the

[9] Accordingly, the number of PAs involving diverging role realizations in Table 1 is strongly underestimated.

[10] Source document ID: AFP_ENG_20080205.0230

[11] Source document ID: APW_ENG_20080206.0766

[12] Source document ID: AFP_ENG_20031015.0353

[13] Source document ID: APW_ENG_20031015.0236

[14] These different configurations are termed *constructional* vs. *lexical licensors* in the SemEval 2010 Task 10 (Ruppenhofer et al., 2010).

[15] Source document ID: AFP_ENG_20011228.0365

[16] Source document ID: APW_ENG_20011228.0572

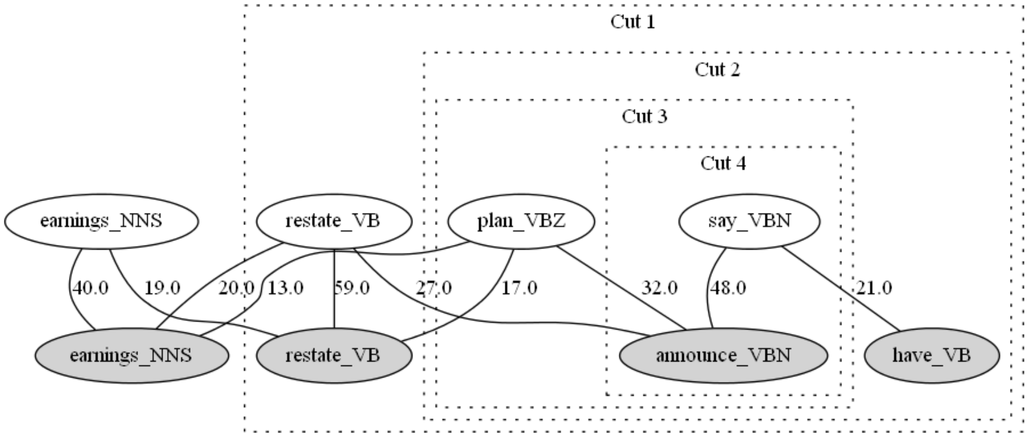

Figure 1: The predicates of two sentences (white: "The company has said it plans to restate its earnings for 2000 through 2002."; gray: "The company had announced in January that it would have to restate earnings (. . .)") from the Microsoft Research Paragraph Corpus are aligned by computing clusters with minimum cuts.

above instances do not provide exhaustive information for grounding all such decisions. A comprehensive model of discourse coherence will need to estimate the argument realization potential of different predicates and roles from larger corpora. But as can be seen from the discussed examples, training a semantic model with suitable discourse features on all predicate argument structures in a large corpus such as ours will provide indicative range of realization decisions.

5 Experiments

This section presents an initial experiment using an unsupervised graph-based clustering method for the task of aligning predicates across comparable texts. We describe the alignment model, two baselines as well as the experimental setting and results.[17]

5.1 Clustering Model

Similarity Measures. We define a number of similarity measures between predicates, which make use of complementary lexical information. One source of information are token-based frequency counts, which we compute over all documents from the AFP and APW sections of Gigaword[18]. Given two lemmatized predicates and their respective PAS, we employ the following four similarity measures: Similarity in WordNet ($\mathbf{sim_{WN}}$) and VerbNet ($\mathbf{sim_{VN}}$), distributional similarity ($\mathbf{sim_{Dist}}$) and bag-of-word similarity of arguments ($\mathbf{sim_{Args}}$). The first three measures are type-based, whereas the latter is token-based.

Graph Representation. The input for graph clustering is a bi-partite graph representation for pairs of texts to be predicate-aligned. In this graph, each node represents a PAS that was assigned during preprocessing (cf. Section 3). Edges are inserted between pairs of predicates that are from two distinct texts. A weight is assigned to each edge by a combination of the introduced similarity measures.

Clustering algorithm. The graph clustering method uses minimum cuts (or *Mincuts*) in order to partition the bipartite text graph into clusters of aligned predicates. Each Mincut operation divides a graph into two disjoint sub-graphs, such that the sum of weights of removed edges will be minimal. As the goal is to induce clusters consisting of pairs of similar predicates, a maximum number of two nodes per cluster is set as stopping criterion. We apply Mincut recursively to the input graph and resulting sub-graphs until we reach the stopping criterion. Figure 1 shows an example of a graph clustered by the Mincut approach.

5.2 Setting

We perform evaluations of the graph-based alignment model (henceforth called **Clustering**) on the

[17] The technicalities of this model, including detailed definitions of the similarity measures, are described elsewhere (manuscript, under submission).

[18] These sections make up 56.6% of documents in Gigaword.

task of inducing predicate alignments across comparable monolingual texts. We evaluate on the manually annotated gold alignments in the test data set described in Section 3.2.

Parameter Tuning. As the graph representation becomes rather inefficient to handle using edges between all predicate pairs, we use the development set of 10 text pairs to estimate a threshold for adding edges. We found the best similarity threshold to be an edge weight of 2.5. Note that the edge weights are calculated as a weighted linear combination of four different similarity measures. Subsequently, we also tune the weighting scheme for similarity measures on the development set. We found the best performing combination of weights to be 0.09, 0.19, 0.48 and 0.24 for $\mathbf{sim_{WN}}$, $\mathbf{sim_{VN}}$, $\mathbf{sim_{Dist}}$ and $\mathbf{sim_{Args}}$, respectively.

Baselines. A simple baseline for this task is to align all predicates whose lemmas are identical (**SameLemma**). As a more sophisticated baseline, we make use of alignment tools commonly used in statistical machine translation (SMT). We train our own word alignment model using the state-of-the-art tool Berkeley Aligner (Liang et al., 2006). As word alignment tools require pairs of sentences as input, we first extract paraphrases for this baseline using a re-implementation of the paraphrase detection system by Wan et al. (2006). In the following sections, we abbreviate this model as **WordAlign**.

5.3 Results

Following Cohn et al. (2008) we measure precision as the number of predicted alignments also annotated in the gold standard divided by the total number of predictions. Recall is measured as the number of correctly predicted *sure* alignments devided by the total number of sure alignments in the gold standard. We subsequently compute the F_1-score as the harmonic mean between precision and recall.

	Precision	Recall	F1
WordAlign	19.7%	15.2%	17.2%
SameLemma	**40.3%**	**60.3%**	**48.3%**
Clustering	**59.7%**	50.7%	**54.8%**

Table 2: Results for all models on our test set; significant improvements ($p<0.005$) over the results given in each previous line are marked in bold face.

Table 2 presents the results for our model and the two baselines. From all four approaches, **WordAlign** performs worst. We identify two main reasons for this: On the one hand, the paraphrase detection does not perform perfectly. Hence, the extracted sentence pairs do not always contain gold alignments. On the other hand, even sentence pairs that contain gold alignments are generally less parallel compared to a typical SMT setting, which makes them harder to align.

We observe that the majority of all sure alignments (60.3%) can be retrieved by applying the **SameLemma** model, yet at a low precision (40.3%). While the **Clustering** model only recalls 50.7% of all cases, it clearly outperforms **SameLemma** in terms of precision (+19.4% points), an important factor for us as we plan to use the alignments in subsequent tasks. With 54.8%, **Clustering** also achieves the best overall F_1-score. We computed statistical significance of result differences with a paired t-test (Cohen, 1995), yielding significance at the 99.5% level for precision and F_1-score.

5.4 Analysis of Results

We perform an analysis of the output of the **Clustering** model on the development set to categorize correct and incorrect alignment decisions.[19] In total, the model missed 13 out of 35 sure alignments (*Type I errors*) and predicted 23 alignments not annotated in the gold standard (*Type II errors*). Six Type I errors (46%) occurred when the lemma of an affected predicate occurred more than once in a text and the model missed the correct link. Vice versa, we find 18 Type II errors (78%) that were made because of a high predicate similarity despite low argument overlap. An example is given in (9).

(9) a. The US alert (...) followed intelligence reports that ...[20]

b. The Foreign Ministry announcement called on Japanese citizens to be cautious ...[21]

While argument overlap itself can be low even for correct alignments, the results clearly indicate that

[19] We decided to leave the test set untouched for further experiments. Due to parameter tuning, the results on the development set also provide us with an upper bound of the proposed model.

[20] Source document ID: AFP_ENG_20101004.0367

[21] Source document ID: APW_ENG_20101004.0207

a better integration of context is necessary: Example (10.a) illustrates a case in which the agent of a warning event is not realized. Here, contextual information is required to correctly align it to one of the warning events in (10.b). This involves inference beyond the local PAS.

(10) a. The US alert (...) is one step down from a full [travel]$_{\text{Arg1}}$ warning [ ]$_{\text{Arg0}}$.[20]

b. Japan has issued a travel alert ...(which) follows similar warnings [from American and British authorities]$_{\text{Arg0}}$. (...) An official said it was highly unusual for [Tokyo]$_{\text{Arg0}}$ to issue such a warning ...[21]

On the positive side, **Clustering** achieves a precision of 61.4% and a recall of 65.7% on the development set. Example (11) shows a correctly aligned PAS pair that involves non-realized arguments:

(11) a. ...the Governing Council has established [a committee]$_{\text{Arg0}}$ to draft [a constitution]$_{\text{Arg1}}$.[22]

b. A .. resolution calls on the Governing Council for elections and the drafting [ ]$_{\text{Arg0}}$ [of a new constitution]$_{\text{Arg1}}$.[23]

In (11.a), the follow-up sentences will refer back to the committee that will draft the new Iraqi constitution, hence the institution has to be introduced in the discourse at this point. In contrast, excerpt (11.b) is the last sentence of a news report. Since it presents a summary, introducing new (omissible) entities at this point would not concord with general coherence principles.

6 Conclusion

In this paper, we presented a novel corpus of comparable texts that provides full discourse contexts for alternative verbalizations. The motivation for the construction of this corpus is to acquire empirical data for studying discourse coherence factors related to argument structure realization. A special phenomenon we are interested in are discourse-related factors that license the omission of argument roles.

Our data set satisfies two conditions that are essential for the purported task: the texts are about the same events and constitute alternative verbalizations. Selected from the Gigaword corpus, the documents pertain to the news domain, and satisfy the further constraint that we have access to the full surrounding discourse context. The constructed corpus could thus be profitable for a range of other tasks that need to investigate factors for knowledge aggregation, such as summarization, or inference in discourse, such as textual entailment.

In total, we derived more than 160,000 document pairs from all pairwise combinations of newswire sources in the English Gigaword Fifth Edition. Using a subset of these pairs, we constructed a development and an evaluation data set with gold alignments that relate predications with (possibly partial) PAS correspondence. We established that the annotation task, while difficult, can be performed with good inter-annotator agreement (κ at 0.86).

We presented first experiments on the task of automatically predicting predicate alignments. This step is essential to gather empirical evidence of different PAS realizations for the same event, given varying discourse contexts. Analysis of the data shows that the aligned predications capture a wide variety of sources and variations of coherence effects, including constructional, lexical and discourse phenomena.

In future work, we will enhance our model by incorporating more refined semantic similarity measures including discourse-based criteria for establishing cross-document alignments. Given that our data set includes sets of aligned documents from several newswire sources, we will explore transitivity constraints across multiple document pairs in order to further enhance the precision of the alignment model. We will then proceed to the ultimate aim of our work: the development of a coherence model for argument structure realization, including the design of an appropriate task and evaluation setting.

Acknowledgements

We are grateful to the Landesgraduiertenförderung Baden-Württemberg for funding within the research initiative “Coherence in language processing” of Heidelberg University. We thank Danny Rehl and Lukas Funk for annotation and Kathrin Spreyer, Tae-Gil Noh and Carina Silberer for helpful discussion.

[22] Source document ID: AFP_ENG_20031015.0353

[23] Source document ID: APW_ENG_20031015.0236.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* Boston, Mass., 2–7 May 2004, pages 113–120.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics,* Toulouse, pages 50–57.

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. 2009. The grec main subject reference generation challenge 2009: overview and evaluation results. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, pages 79–87.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstration Volume*, pages 33–36, Beijing, China, August. Coling 2010 Organizing Committee.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China, August. Coling 2010 Organizing Committee.

Chris Brockett. 2007. *Aligning the RTE 2006 Corpus*. Microsoft Research.

Aoife Cahill and Arndt Riester. 2009. Incorporating information status into generation ranking. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 817–825, Suntec, Singapore, August. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Paul R. Cohen. 1995. *Empirical methods for artificial intelligence*. MIT Press, Cambridge, MA, USA.

Trevor Cohn, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing Corpora for Development and Evaluation of Paraphrase Systems. 34(4).

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing*.

Katja Filippova and Michael Strube. 2007. Generating constituent order in German clauses. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,* Prague, Czech Republic, 23–30 June 2007, pages 320–327.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16.3:235–250.

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

Shudong Huang, David Graff, and George Doddington. 2002. *Multiple-Translation Chinese Corpus*. Linguistic Data Consortium, Philadelphia.

Philipp Koehn, 2005. *Europarl: A parallel corpus for statistical machine translation*, volume 5, pages 79–86.

Percy Liang, Benjamin Taskar, and Dan Klein. 2006. Alignment by agreement. In *North American Association for Computational Linguistics (NAACL)*, pages 104–111.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory. Toward a functional theory of text organization. *Text*, 8(3):243–281.

Kathleen R. McKeown and Dragomir Radev. 1995. Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval,* Seattle, Wash., 9–13 July 1995, pages 74–82. Reprinted in *Advances in Automatic Text Summarization*, Mani, I. and Maybury, M.T. (Eds.), Cambridge, Mass.: MIT Press, 1999, pp.381-389.

Adam Meyers, Ruth Reeves, and Catherine Macleod. 2008. *NomBank v1.0*. Linguistic Data Consortium, Philadelphia.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Robert Parker, David Graff, Jumbo Kong, Ke Chen, and Kazuaki Maeda. 2011. *English Gigaword Fifth Edition*. Linguistic Data Consortium, Philadelphia.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2010. SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluations*, pages 45–50, Uppsala, Sweden, July.

Ivan Titov and Mikhail Kozhevnikov. 2010. Bootstrapping semantic analyzers from non-contradictory texts.

In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11–16 July 2010, pages 958–967.

Stephen Wan, Mark Dras, Robert Dale, and Cecile Paris. 2006. Using dependency-based features to take the ”Para-farce” out of paraphrase. In *Proceedings of the Australasian Language Technology Workshop*, pages 131–138.

Sander Wubben, Antal van den Bosch, Emiel Krahmer, and Erwin Marsi. 2009. Clustering and matching headlines for automatic paraphrase acquisition. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 122–125, Athens, Greece, March. Association for Computational Linguistics.