Annotating Relation Mentions in Tabloid Press

Hong Li, Sebastian Krause, Feiyu Xu, Hans Uszkoreit, Robert Hummel, Veselina Mironova

Language Technology Lab
German Research Center for Artificial Intelligence (DFKI)
Alt-Moabit 91c, 10559 Berlin, Germany
dare@dfki.de

Abstract

This paper presents a new resource for the training and evaluation needed by relation extraction experiments. The corpus consists of annotations of mentions for three semantic relations: *marriage*, *parent–child*, *siblings*, selected from the domain of biographic facts about persons and their social relationships. The corpus contains more than one hundred news articles from Tabloid Press. In the current corpus, we only consider the relation mentions occurring in the individual sentences. We provide multi-level annotations which specify the marked facts from relation, argument, entity, down to the token level, thus allowing for detailed analysis of linguistic phenomena and their interactions. A generic markup tool Recon developed at the DFKI LT lab has been utilised for the annotation task. The corpus has been annotated by two human experts, supported by additional conflict resolution conducted by a third expert. As shown in the evaluation, the annotation is of high quality as proved by the stated inter-annotator agreements both on sentence level and on relation-mention level. The current corpus is already in active use in our research for evaluation of the relation extraction performance of our automatically learned extraction patterns.

Keywords: Corpus Annotation, Relation Extraction, Semantic Relations, People Domain, Tabloid Press

1. Introduction

In this paper we present a new resource for performing experiments in the area of text analysis and information extraction, namely a corpus consisting of more than one hundred documents annotated with mentions of several semantic relations. The targeted relations were selected from the domain of biographic facts about persons and their family relatives. The provided annotation specifies the marked facts down to the token level, it thereby allows for detailed analysis of language phenomena.

This resource is intended to foster research in the task of detecting mentions of semantic relations (relation extraction; RE) in texts and extracting factual knowledge from them. The corpus is particularly useful to evaluate and compare the output of competing RE systems. Such comparison attempts are commonly hard to perform because of the lack of freely available gold-standard corpora. The typical fallback solutions for comparing systems then have to rely on ad-hoc methods, such as manual investigation of output samples or fuzzy matching of data against some database of known facts, thus, the results of such a comparison are of limited significance. An additional benefit of our resource is that it enables the training of classic supervised RE systems.

The corpus was annotated using the markup tool Recon (Li et al., 2012) by two human experts with additional conflict resolution performed by a third expert. The annotation is of high quality and is already in active use in our RE research, some of which is described in Section 5. Other examples include the work we described in (Xu et al., 2010) and (Krause et al., 2012), where we utilized preliminary versions of this corpus. We hope that the publication of this resource will encourage other researchers to work on the interesting problems of this research field.

2. Related Work

There has been some work on preparing evaluation corpora for relation extraction in the past. An early example includes the Message Understanding Conference series in the eighties and nineties, which prepared different text collections with annotations in order to use them for information extraction competitions, e. g., the MUC 6 corpus (Chinchor and Sundheim, 2003) dealing with business related information. Another prominent example is the Automatic Content Extraction program, which led to the development of evaluation corpora, like the ACE 2005 corpus (Walker et al., 2006).

These existing corpora are unfortunately not appropriate for certain application scenarios. For example, the ACE 2005 corpus contains relatively few mentions per relation, making it hard to evaluate the impact of filtering strategies for pattern-based RE approaches.

Recently, corpus annotation efforts focused on automatic high-precision annotation of corpora, in particular with mentions of entities and their disambiguation. Examples are the Wikilinks corpus (Singh et al., 2012) and the "Freebase Annotation of the ClueWeb Corpora" (Gabrilovich et al., 2013), each of them containing millions of marked entity mentions.

3. Annotated Semantic Relations

The annotation goal was to create a corpus suitable for utilization in different relation extraction scenarios, for example to evaluate learned extraction patterns or to train statistical classifiers. We decided to annotate three semantic relations from the domain of biographic facts about people and their family relatives, because previous work (Krause et al., 2012) showed that these relations have interesting properties and that they are frequently mentioned in certain genres of news articles. The selected relations are listed in Table 1.

Relation	Argument	Type	Required	Size
marriage	SPOUSE	person	x	$= 2$
	FROM	date		≤ 1
	TO	date		≤ 1
	CEREMONY	location		≤ 1
parent–child	CHILD	person	x	≥ 1
	PARENT	person	x	$\{1, 2\}$
siblings	SIBLING	person	x	≥ 2

Table 1: Definition of the annotated relations.

We annotate relation mentions on the basis of individual sentences. For a relation mention to be considered valid, it must consist of at least two arguments. An argument is in turn a mention of a specific concept type, such as person, location or date. Example 1 shows two sentences from the corpus and the corresponding relation-mention annotation (underlined).

Example 1

a) On July 14 [...] Molly walked down the aisle of St. Timothy's Church on her father's arm to wed her fiancé, Keith Ormrod, 28. [1]

Semantic Relation: *marriage*				
Argument	SPOUSE	SPOUSE	FROM	CEREMONY
Concept Mention	*her (⇝ Molly)*	*Keith Ormrod*	*July 14*	*St. Timothy's Church*

b) [...] as he takes a seat beside his wife, Dorothea, who dandles their 6-month-old, Jacob. [2]

Semantic Relation: *marriage*		
Argument	SPOUSE	SPOUSE
Concept Mention	*his (⇝ he)*	*Dorothea*

Semantic Relation: *parent–child*			
Argument	PARENT	PARENT	CHILD
Concept Mention	*his (⇝ he)*	*Dorothea*	*Jacob*

4. Relation Mention Annotation

4.1. Corpus: *"Celebrity"*

Xu et al. (2010) and Krause et al. (2012) conducted experiments on a celebrity-gossip domain, using a subset of a collection of PEOPLE-magazine articles from the years 2001–2008. We selected the 150 longest documents from the same article basis to build our RE corpus, dubbed "Celebrity".

After duplicate removal 142 documents with 364,400 words/2.1 MB remained. To speed up the annotation process, we preprocessed the corpus with the entity recognizers OpenCalais[3] and SProUT [4] (Drozdzynski et al., 2004) to recognize mentions of persons, organizations, locations and date expressions. Approximately 30k mentions of concepts were detected. Because our annotation effort is focused on relations instances mentioned within individual sentences, we additionally segmented the sentences using the English tagger included in Stanford CoreNLP[5] and obtained about 25k sentences.

4.2. Annotation Tool: Recon

We used the annotation tool Recon (Li et al., 2012) for marking mentions of concepts and relation mentions in documents. Recon is a general and flexible annotation tool for annotating n-ary relations among text elements. Not needing any relation definitions beforehand, the annotator can start right away with marking arbitrary text spans as concept mentions and can assemble these later together with argument-role labels to create relation mentions. Figure 1 exemplifies how the *marriage* relation mention from Example 1a) is annotated in Recon.

4.3. Single-Expert Annotation

In the first phase of the annotation process we had two independently working experts, who manually annotated the whole corpus with Recon.

Annotation Tasks

The tasks of the annotators were:

1. Identification of sentences with relation mentions
2. For sentences from 1): Verification of automatic concept-mention annotation of the NLP tools, including information about coreference resolution for personal pronouns
3. For sentences from 1): Annotation of relation mentions (i. e., assignment of argument roles to concept mentions)

Note that we discarded the automatic concept-mention annotation for sentences without any relation mentions, due to the effort required to manually verify it. For the same reason, we did not perform full resolution of coreferring expressions, but register for mentions of persons the feature "name", whose value is set to the canonical name of the referenced entity, if available in the document.

In some cases, an entity being part of a relation mention is referenced to by several concept mentions within a sentence. For example, in Example 1a), both "*Molly*" and "*her*" could be annotated as the SPOUSE argument of the *marriage*-relation mention. To have the experts annotate in a consistent way for such cases, we define the "nearest arguments" principle, stating that the nearest concept mentions should serve as the arguments of the relation mention. Therefore, "*her*" in Example 1a) and "*his*" in Example 1b) are marked as relation arguments, instead of "*Molly*"/"*he*".

[1] http://www.people.com/people/archive/article/0,,20135040,00.html

[2] http://www.people.com/people/archive/article/0,,20138581,00.html

[3] http://www.opencalais.com/

[4] http://sprout.dfki.de/

[5] http://nlp.stanford.edu/software/corenlp.shtml

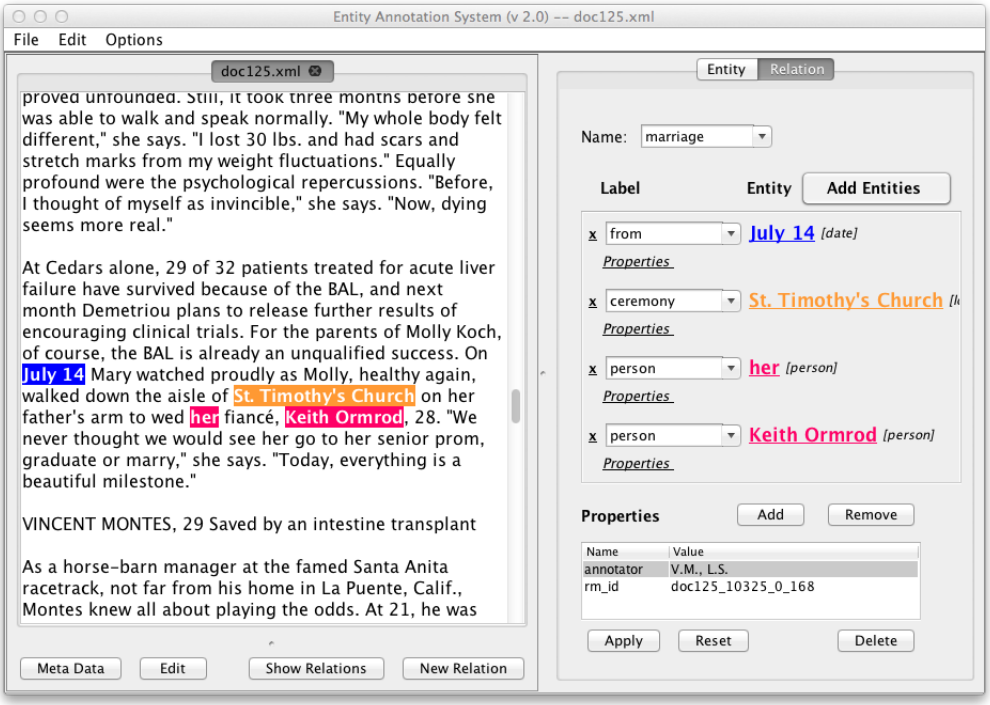

Figure 1: Annotation example in Recon, cf. Example 1a).

Annotation Results

Table 2 presents statistics of the annotation of the two annotators (referenced by A and B).

Annotator	Sentences w/ Relation Mention	Concept Mentions	Relation Mentions
A	908	4003	1125
B	872	3542	1081
$A \cup B$	971	4090	1318

Table 2: Overview of the annotation results by two annotators A and B

Of the approximately 25,000 sentences of the corpus, only 4% have been annotated with relation mentions by at least one of the experts. About 80% of these sentences contain exactly one relation mention, 15% have two mentions and the remaining 5% have three or more relation mentions. The average number of concept mentions in the sentences with relation mentions is 4.2.

4.4. Inter-annotator Agreement

After the initial annotation, we calculated the agreement between the two experts both on sentence level and on relation-mention level.

Agreement on Sentence Level

The sentence-level agreement is evaluated by reducing the complexity of the annotation to the binary choice whether a given sentence of the corpus contains a mention of a given target relation. This abstraction allows to calculate standard inter-annotator agreement measures, depicted in Table 3.

The agreement between the annotators is very good for all three relations. This indicates that relations between persons are commonly expressed in a clear and relatively objective way. Note that the two annotators have a higher agreement for the relation *marriage* than for the other two relations.

	marriage	*parent–child*	*siblings*	micro-avg.
Pos. Agreem.	0.9358	0.8745	0.8545	0.8926
Cohen's κ	0.9348	0.8721	0.8532	0.8910
PCC	0.9349	0.8730	0.8546	0.8554

Table 3: Inter-annotator agreement measures on sentence level. PCC is short for Pearson correlation coefficient.

Agreement on Relation-Mention Level

To measure the agreement of the actual relation mention annotation, we use the *agr* metric described by Wiebe et al. (2005):

$$agr(A||B) = \frac{\text{\# of relation mentions annotated by } A \text{ and } B}{\text{\# of relation mentions annotated by } A} \quad (1)$$

Table 4 presents for all relations the number of mentions annotated by $A \cap B$, A/B, B/A, and also the agreement metrics. Consistent with the agreement on the sentence level, the agreement for the relation *marriage* is higher than for the other two relations. This might be caused by the potentially higher number of persons in a sentence which are related by *parent–child*/*siblings* than for *marriage*, thus having a higher potential of disagreement. Another reason might be that the annotators disagree about specific details of the relation definitions.

Consider Example 2, where the annotators disagreed about the role of "*Rosemary*", i. e., whether this sentence is suf-

	marriage	*parent–child*	*siblings*	micro-avg.
$A \cap B$	342	368	176	–
A/B	47	139	53	–
B/A	61	101	31	–
$agr(A\|\|B)$	0.8792	0.7258	0.7686	0.7876
$agr(B\|\|A)$	0.8486	0.7846	0.8502	0.8211
F1	0.8636	0.7541	0.8073	0.8040

Table 4: Inter-annotator agreement on relation-mention level.
$A \cap B$: Mentions annotated by both annotators using the same required arguments.
A/B: Mentions annotated exclusively by A.
B/A: Mentions annotated exclusively by B.

ficient evidence to conclude that she is the mother of the mentioned children of "*Mike*". To resolve this problem, we included step-parents into the *parent–child* definition for the annotation-merging step.

Example 2 *"..." says Mike, a laid-off manager who lives with wife Rosemary, 41, and sons Stanley, 19, Tony, 15, and Chris in Santee, Calif.*

4.5. Annotation Merging

After the two experts finished their annotation runs, we performed a combination of automatic and manual conflict resolution methods. At first, we merged the annotated concept

Annotator	Consistent	Conflict		Total
		no overlap	disagree	
A	3,455	530	18	4003
B		70	17	3542

Table 5: Overview of the annotated concept mentions in sentence with relation mentions

mentions in the sentences with marked relation mentions. The major fraction of the concept mentions were marked in exactly the same way by both annotators. Table 5 shows the overview of the annotated concept mentions. Conflict cases here included mentions only annotated by one of the experts, or disagreement in the exact extent of the mention, as illustrated by Example 3.

Example 3 *"...," says Dixie Chick Natalie Maines, explaining ...*

Here, the two types of lines represent the different mark-up of the two annotators. Conflict resolution was performed by using the longest entity mention of two conflicting ones. In the end, we obtained approximately 4k concept mentions in the 971 sentences with relation mentions.
The last step was to merge the annotated relation mentions. The second and third row of Table 4 list the number of relation mentions with complete disagreement between the annotators. The first row of the table states the number of relation mentions for which the annotators agreed on at least the required arguments. For only eleven of them the experts marked different optional arguments.

All of these approximately 400 relation mentions with disagreement in either required or optional arguments were again checked by a third human expert (to which we refer by C), Table 6 shows the results of this process.

	Mentions from A	B
C agrees	214	162
C disagrees	36	42
$agr(\cdot\|\|C)$	0.856	0.794

Table 6: Conflict resolution on relation mentions by third annotator C.

Using all 875 conflict-free relation mentions annotated both by A and B and all the correct mentions as judged by C and removing projections of mentions, we finally obtained 1,220 relation mentions. Table 7 shows the final statistics of the annotated corpus.

5. Evaluation of Relation Extraction

To illustrate the usefulness of our corpus for RE evaluation, we applied all of the learned extraction patterns from (Krause et al., 2012) to the new corpus. Table 8 compares the statistics of this rule application to the ones from (Moro et al., 2013), i. e., the application to the English version of the Gigaword corpus (Parker, 2011). Note that we did not apply any filtering to the candidate RE patterns here, so the precision values are relatively low. Also note that the evaluation of precision and recall values for the Gigaword part required manual evaluation of a sample of the matched mentions and the corpus, resulting both in non-repeatability and a low significance of the values.
While our corpus has advantages in these aspects, unfortunately it lacks the large-scale property of the Gigaword corpus, having, e. g., only one tenth of the number of applied patterns the Gigaword corpus has.

		Rule Application		**Evaluation**	
		Applied Rules	**Extracted Mentions**	**Precision**	**Recall**
marriage	G	5,337	92,780	11.60%	38.61%
	C	504	1,182	16.49%	50.96%
parent–child	G	2,792	93,800	13.20%	38.23%
	C	358	1,196	17.89%	40.76%
siblings	G	1,856	59,465	5.60%	26.08%
	C	204	735	4.89%	18.36%

Table 8: Comparison of extraction-pattern application to our corpus (C) and the Gigaword corpus (G).

6. Conclusion and Future Work

In this paper we present a published corpus for the training and evaluation of RE systems, together with a detailed description of the methodology we followed to prepare the corpus. Despite the followed free-text annotation paradigm, the resulting annotation is of high quality, as proved by the stated inter-annotator agreements.

Documents	Sentences	Sentences with Relation Mentions	Concept Mentions as Rel. Arguments	Relation Mentions			
				marriage	*parent–child*	*siblings*	**Total**
142	25,065	967	2,506	421	550	249	1,220

Table 7: Final statistics of the published corpus.

In our future work we plan on the one hand to annotate this and other corpora with relations from different domains (e. g., business related) and also want to employ a hierarchy for the relation arguments (e. g., SPOUSE → WIFE/HUSBAND). On the other hand, we want to annotate the corpora with more detailed concept information, including complete resolution of coreferring expressions, both in- and cross-document-wise.

7. Acknowledgements

This research was partially supported by the German Federal Ministry of Education and Research (BMBF) through the project Deependance (contract 01IW11003) and by Google through a Focused Research Award for the project LUcKY granted in July 2013.

8. References

Nancy Chinchor and Beth Sundheim. 2003. Message understanding conference (MUC) 6. Linguistic Data Consortium, Philadelphia.

Witold Drozdzynski, Hans-Ulrich Krieger, Jakub Piskorski, Ulrich Schäfer, and Feiyu Xu. 2004. Shallow processing with unification and typed feature structures — foundations and applications. Künstliche Intelligenz, 1:17–23.

Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. 2013. Facc1: Freebase annotation of ClueWeb corpora, version 1.

Sebastian Krause, Hong Li, Hans Uszkoreit, and Feiyu Xu. 2012. Large-scale learning of relation-extraction rules with distant supervision from the web. In Proceedings of the 11th International Semantic Web Conference. Springer, 11.

Hong Li, Xiwen Cheng, Kristina Adson, Tal Kirshboim, and Feiyu Xu. 2012. Annotating opinions in german political news. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, may. European Language Resources Association (ELRA).

Andrea Moro, Hong Li, Sebastian Krause, Feiyu Xu, Roberto Navigli, and Hans Uszkoreit. 2013. Semantic rule filtering for web-scale relation extraction. In 12th International Semantic Web Conference, Sydney, Australia, October.

Robert Parker. 2011. English gigaword fifth edition. Linguistic Data Consortium, Philadelphia.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2012. Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia. Technical Report UM-CS-2012-015.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. Linguistic Data Consortium, Philadelphia, February.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. volume 39(2/3), page 165210.

Feiyu Xu, Hans Uszkoreit, Sebastian Krause, and Hong Li. 2010. Boosting relation extraction with limited closed-world knowledge. In Coling 2010: Posters, pages 1354–1362, Beijing, China, August. Coling 2010 Organizing Committee.