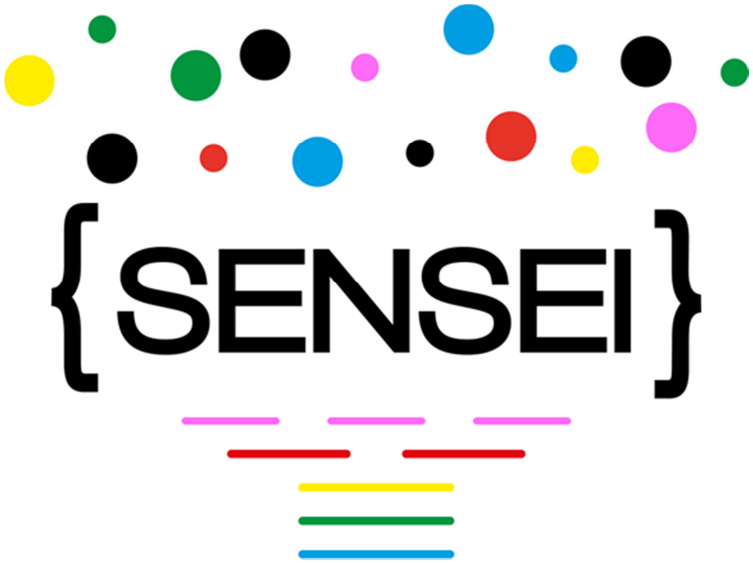

D4.1 – Discourse Descriptions of Conversations

Document Number	D4.1
Document Title	Discourse descriptions of conversations
Version	2.0
Status	Draft
Work Package	WP4
Deliverable Type	Report
Contractual Date of Delivery	31.10.2014
Actual Date of Delivery	31/10/2014
Responsible Unit	UESSEX
Keyword List	discourse parsing, event/temporal structure, argumentation structure, intra/inter document coreference
Dissemination level	PU

Editor

Mijail Kabadjov (UESSEX)

Evgeny. A. Stepanov (UNITN)

Contributors

Adam Funk (USFD)

Benoit Favre (AMU)

Mijail Kabadjov (UESSEX)

Massimo Poesio (UESSEX)

Evgeny. A. Stepanov (UNITN)

SENSEI Coordinator

Prof. Giuseppe Riccardi

Department of Information Engineering and Computer Science

University of Trento, Italy

riccardi@disi.unitn.it

Document change record

Version	Date	Status	Author (Unit)	Description
0.1	28/07/2014	Draft	M. Poesio, M. Kabadjov (UESSEX)	Table of Contents
0.2	28/08/2014	Draft	E. Stepanov, G. Riccardi (UNITN)	Section 1 added
0.3	6/09/2014	Draft	M. Poesio, M. Kabadjov (UESSEX)	Section 3 added
0.4	9/9/2014	Draft	Benoit Favre (AMU)	AMU's contribution
0.5	9/9/2014	Draft	M. Kabadjov (UESSEX)	Moved AMU's contribution to section 1.4.
0.6	21/9/2014	Draft	A. Funk (USFD), M. Kabadjov, M. Poesio (UESSEX)	USFD Provided text for section 2.1. UESSEX added the executive summary
0.7	22/9/2014	Draft	M. Poesio (UESSEX)	UESSEX added Introduction and Index
0.8	30/09/2014	Draft	M. Poesio (USSEX) E. Stepanov, G. Riccardi (UNITN)	Added section 1. Amended and extended section 2
0.9	3/10/2014 6/10/2014	Draft	M. Kabadjov (UESSEX) E. Stepanov (UNITN)	Added section 3.2. provided by A. Funk (USFD), amended and extended section 4. Additions and references.
1.0	6/10/2014	Draft	E. Chiarani (UNITN)	Quality check completed
1.1	09/10/2014	Draft	M. Kabadjov (UESSEX)	Added keywords, conclusions, and a few other revisions.
1.2	17/10/2014	Draft	F. Bechet (AMU) M. Kabadjov (UESSEX) A. Funk (USFD)	Internal Review Addressed comments from internal review. Provided revision for section 3.
2.0	20/10/2014	Final	E. Chiarani, G. Riccardi (UNITN)	Final version for submission

Executive summary

In this deliverable we present the progress on the discourse analysis methods developed within the project in Year 1. Three lines of work were carried out: on discourse parsing of spoken conversations in French and Italian, on extracting event structure from English texts, and on intra-document coreference in social media in Italian and English.

The document is organised as follows: in Section 1, progress on discourse parsing for conversations is presented, next, work on event extraction and temporal structure from conversation is discussed, and finally, progress on intra-document coreference resolution for conversations and social media is described.

Table of Content

1. Introduction

The objective of WP4 is to develop tools supporting automated discourse analysis of conversations—specifically, discourse parsing, event/temporal structure, argumentation structure, and intra/inter document coreference—in the two domains (social media conversations and call center conversations) and three languages (English, French, and Italian) of the project. A particular focus of the research is to investigate the performance of techniques developed for the most extensively studied forms of language use (news) in these new domains, and develop methods for adapting such methods. The objectives for Year 1 were to develop the first release of these tools; a strong emphasis was placed on domain adaptation methods, and on the creation of appropriate resources for carrying out this work when none were already available.

In Deliverable D4.1 we present the progress on this line of research. It provides details on the tools used for the three tasks of work package 4, the experiments carried out on developing and adapting these tools to the domains of interest to the project, the current challenges faced by the various teams working on the tasks and the next steps to pursue as the project moves into its second year.

One of the foci of Task 4.1 (discourse parsing of conversations) is to develop methods to adapt discourse parsing to different domains and conversational styles. During SENSEI Year 1, the discourse parsing pipeline of Stepanov & Riccardi (2013) was tested for cross-domain and genre generalization (Stepanov, 2014). As a first step toward discourse parsing of spoken conversations, a sentence-like unit tagger based on a CRFapproach was trained on call-center conversations from RATP-DECODA.

Similarly, in the work on temporal and event structure, an event detection component developed in ARCOMEM was adapted to the new domain. Two approaches were combined: a top-down, template filling approach, and a bottom-up, verbal relations-driven approach.

In the work on intra-document coreference, an existing resource (the Blog subcorpus of the LiveMemories Anaphora corpus of intra-document coreference in Italian) was used to adapt the latest, state of the art version of the BART toolkit on social media data. A new dataset for English was also created, collecting data from the Guardian via the SENSEI tools developed by WebSays and annotating them using the same guidelines.

The document is organised as follows: in Section 2, progress on discourse parsing for conversations is presented. Next, work on event extraction and temporal structure from conversation is discussed. Finally, progress on intra-document coreference resolution for conversations and social media is described.

2. Discourse parsing for conversations

SENSEI Task 4.1 focuses on adapting and evaluating discourse parsing methods to a range of conversational styles and to novel domains. We adopt Penn Discourse Treebank (PDTB) (Prasad et al., 2008) style discourse parsing, which is identified by its non-hierarchical binary view on discourse relations: Argument 1 (Arg1) and Argument 2 (Arg2), where Arg2 is syntactically attached to a discourse connective. Thus, a discourse relation is a triplet of a connective and its two arguments. In the literature (Lin et al., 2012; Stepanov and Riccardi, 2013, 2014) PDTB-style discourse parsing is partitioned into *discourse relation detection, argument position classification, argument span extraction, and relation sense classification.*

A discourse connective is a member of a closed class (e.g. 100 connectives for PDTB). A discourse relation signaled by a discourse connective is an explicit discourse relation. However, a discourse relation can hold also without the presence of a connective. In such implicit discourse relations, a connective can be inserted, but it is left implicit. In case a connective cannot be inserted while there is a discourse relation between sentences, the discourse relation is said to be *alternatively lexicalized*. There are also other relations annotated in PDTB - Entity Relations and No Relations.

For the explicit discourse relations (i.e., signalled by a connective), *discourse relation detection* is cast as classification of connectives as discourse and non-discourse. *Argument position classification*, on the other hand, involves detection of the location of *Arg1* with respect to *Arg2*, that is to detect whether a relation is inter- or intra- sentential. *Argument span extraction* is the extraction (labelling) of text segments that belong to each of these arguments. Finally, relation sense classification is the annotation of relations with the senses from the sense hierarchy defined for the domain (thus corpus). After explicit relations are identified, a piece of text is inspected for any implicit discourse and non-discourse relations that might hold between two adjacent sentences.

2.1 PDTB-Style Discourse Parsing Algorithms

While discourse relation detection is unambiguous in its algorithmic application, since it is a first step, the other sub-processes allow variability.

Relation sense classification, for instance, for explicit relations can be applied right after the *discourse connection classification* step, considering only the connective for classification, or after the *argument span extraction* step, also considering contents of the argument spans.

Argument span extraction allows variability as well. Since arguments of explicit discourse relations can appear in the same sentence or in different ones (i.e., relations can be intra- or inter-sentential), in the literature there are two approaches to the task. In the first approach the parser decision is not conditioned on whether the relation is intra- or inter-sentential (e.g., (Ghosh et al., 2011)). In the second approach relations are parsed separately for each class (e.g., (Lin et al., 2012; Xu et al., 2012; Stepanov and Riccardi, 2013, 2014)). In the former approach argument span extraction is applied right after discourse connective detection, while the latter approach also requires *argument position classification* step.

Stepanov and Riccardi (2013) have compared the two approaches and demonstrated that on PDTB explicit discourse relations the latter approach outperforms the former one. Their work follows the approach of (Ghosh et al., 2011) and the decision on argument spans is made on token-level, and the problem is cast as sequence labelling using conditional random fields (CRFs) (Lafferty et al., 2001).

The discourse relation parsing pipeline of Stepanov and Riccardi (2013) is presented on Fig. 1. Besides the sub-tasks already presented, it additionally makes use of heuristics for the argument span extraction of inter-sentential discourse relations. (1) Based on the observations in the PDTB, *immediately previous sentence* heuristics selects the sentence immediately preceding the sentence containing a discourse connective classified as an inter-sentential one, as a candidate for *Arg1*. Similarly for Arg2, whole sentence containing the connective is selected as a candidate.

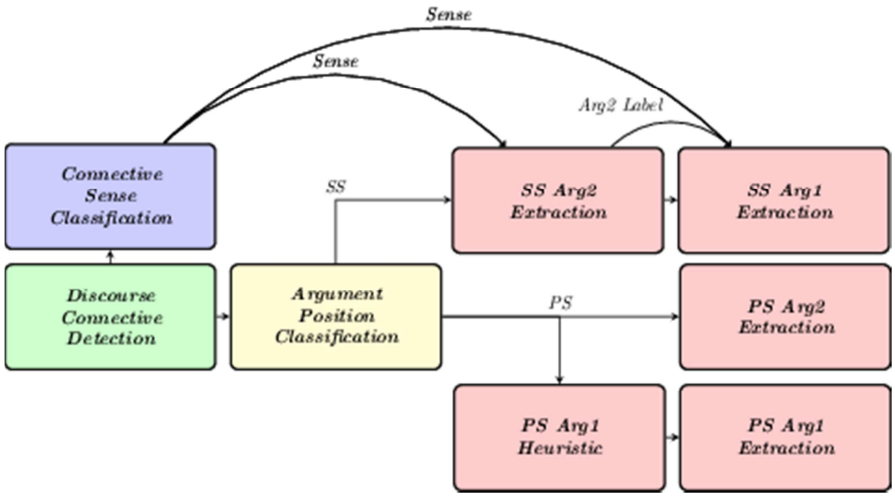

Figure 1: Discourse Parsing Architecture for PDTB explicit discourse relations. CRF Argument Span Extraction models are in bold.

2.1.1. Discourse Parsing Subtask Performances on PDTB

The addDiscourse tool (Pitler and Nenkova, 2009) is used for discourse connection detection and relation sense classification on PDTB. The authors report high accuracy on the tasks on a 10-fold cross-validation setting on PDTB sections 02-22: F-measures of 94.19 and 94.15 for discourse connective detection and relation sense classification into 4 top senses respectively.

For the task of argument position classification, Stepanov and Riccardi (2013) train boostexter model on PDTB sections 02-22, and test on sections 23-24. Since English discourse connectives have a strong preference on the locations of their Arg1 with respect to their grammatical category -- coordinating, subordinating, or discourse adverbial -- and a position in a sentence -- initial or medial; the task has very high baseline and even higher machine learning performance. Stepanov and Riccardi (2013) report F-measure of 98.12 using gold features and 97.81 using automatic features extracted from parse trees produced by Stanford Parser (Klein and Manning, 2003).

Table 1 shows performances on the argument span extraction task with gold features and the features extracted from automatic parse trees. The results reported for each of the arguments (Arg2 and Arg1) for intra- and inter-sentential relations (SS and PS respectively) and jointly (ALL).

Table 1: performances on the argument span extraction task

	Arg2			*Arg1*		
	P	**R**	**F1**	**P**	**R**	**F1**
GOLD						
SS	90.36	87.49	88.90	70.27	66.67	68.42
PS	79.01	77.10	78.04	46.23	36.61	40.86
ALL	85.93	83.45	84.67	61.94	54.98	58.25
AUTO						
SS	86.83	85.14	85.98	64.26	63.01	63.63
PS	75.00	73.67	74.33	37.66	37.00	37.33
ALL	82.24	80.69	81.46	53.93	52.92	53.42

The performance of the PDTB argument span extraction models on the test set with 'Gold' and 'Automatic' sentence splitting, tokenization, and syntactic features. The results are reported together with the error propagation from argument position classification for intra-sentential (SS), inter-sentential (PS) models and joined results (ALL) as precision (P), recall (R) and F-measure (F1). (From Stepanov and Riccardi, 2014)

The general trend in the literature is that the argument span extraction for *Arg1* has lower performance than for *Arg2*, which is expected since *Arg2* position is signalled by a discourse connective. Additionally, Previous Sentence *Arg1* model performance is much lower than that of the other models due to the fact that it only considers immediately previous sentence; which covers only 71.7% of the inter-sentential relations. In the next subsections, these models are evaluated in terms of their cross-domain generalization.

2.2 Domain Adaptation of Discourse Relation Parsing

The goal of this Section is to overview the existing cross-domain studies on the PDTB-style discourse parsing subtasks and complement them with additional experimentation. Out of the four subtasks identified in the previous Section Discourse Connective Detection and Relation Sense Classification have received attention and a PDTB-BioDRB cross-domain studies were carried out. Within SENSEI project we have performed cross-domain evaluation of the other subtasks – Argument Position Classification and Argument Span Extraction. The results of these experiments are published in Stepanov and Riccardi (2014).

2.2.1. Cross-Domain Discourse Connective Detection

The difference between the two corpora (PDTB and BioDRB) with respect to discourse connectives is that in case of PDTB the annotated connectives belong to one of the three syntactic classes: subordinating conjunctions (e.g. *because*), coordinating conjunctions (e.g. *but*), and discourse adverbials (e.g. *however*), while BioDRB is also annotated for a forth syntactic class – subordinators (e.g. *by*).

There are 100 unique connective types in PDTB (after connectives like *1 year after* are stemmed to *after*) in 18,459 explicit discourse relations. Whereas in BioDRB there are 123 unique connective types in 2,636 relations. According to the discourse connective analysis in (Ramesh et al., 2012), the subordinators comprise 33% of all connective types in BioDRB. Additionally, 11% of connective types in common syntactic classes that occur in BioDRB do not occur in PDTB; e.g. *In summary*, *as a consequence*. Thus, only 56% of connective types of BioDRB are common to both corpora. While in-domain discourse connective detection has good performance (Ramesh and Yu, 2010), this difference makes the cross-domain identification of discourse connectives a hard task, which is exemplified by experiments in (Ramesh and Yu, 2010) (F1 = 0.55).

Table 2 presents the results on in-domain and cross-domain discourse connective detection performance.

Table 2: In-domain and cross-domain discourse connective detection performance, reported as F-measure.

Classifier	Gold Features	Automatic Features
PDTB-PDTB		
MaxEnt (Pitler & Nenkova, 2009)	92.75	91.00
MaxEnt (Lin et al., 2012)	95.76	93.62
CRF (Ramesh & Yu, 2010)	84.00	--
BioDRB-BioDRB		
CRF (Ramesh & Yu, 2010)	--	69.00
CRF (Ramesh et al. 2012)	--	75.70
MaxEnt (Faiz & Mercer, 2013)	--	82.36
PDTB-BioDRB		
CRF (Ramesh & Yu, 2010)	--	55.00
CRF (Ramesh et al., 2012)	--	59.20

2.2.2. Cross-Domain Relation Sense Classification

Discourse relations are annotated using a hierarchy of senses: even though the organization of senses and the number of levels are different between the corpora, the most general top level senses are mapped to the PDTB top level senses: *Comparison*, *Contingency*, *Expansion*, and *Temporal* (Prasad et al., 2011).

With respect to relation sense classification, the connective surface provides already high baselines (Prasad et al., 2011). However, cross-domain sense classification experiments indicate that there are significant differences in the semantic usage of connectives between two domains, since the performance of the classifier trained on PDTB does not generalize well to BioDRB (F1 = 0.57).

Table 3 presents the in-domain and cross-domain relation sense classification performance in terms of F-measure. The reported results are for discourse connective token only classifiers and classifiers using other features, such as syntactic and positional.

Table 3: In-domain and cross-domain relation sense classification performance, reported as F-measure.

Classifier	Token Only	Complex
PDTB-PDTB		
N.Bayes (Pitler & Nenkova, 2009)	93.67	94.15
SLIPPER (Prasad et al., 2011)	90.10	--
BioDRB-BioDRB		
SLIPPER (Prasad et al., 2011)	90.90	--
PDTB-BioDRB		
SLIPPER (Prasad et al., 2011)	57.00	--

To sum up, the corpora differences with respect to discourse connective usage affect the cross-domain generalization of connective detection and sense classification tasks negatively.

2.2.3. Cross-Domain Argument Position Classification

For Argument Position Classification the unigram BoosTexter (Schapire and Singer, 2000) model with 100 iterations is trained on PDTB sections 02-22 and tested on sections 23-24. Similar to the previously published results, it has a high performance: F1 = 98.12. The features are connective surface string, POS-tags, and IOB-chains. The results obtained with automatic sentence splitting, tokenization, and syntactic parsing using Stanford Parser (Klein and Manning, 2003) are also high F1 = 97.81 (see Table 4).

Since, unlike PTB for PDTB, for BioDRB there is no manual sentence splitting, tokenization, and syntactic tree annotation; the precise cross-domain evaluation of Argument Span Extraction step is not possible. In order to evaluate cross-domain argument position classification we evaluate classifier decisions against automatic sentence splitting using Stanford Parser (Klein and Manning, 2003) on whole of BioDRB.

The trained BoosTexter model has a high in-domain performance of 97.81. On BioDRB its performance is 95.26, which is still high (see Table 4). Thus, we can conclude that argument position classification generalizes well cross-domain, and that it is little affected by the presence of ‘subordinators’ that were not annotated in PDTB.

Table 4: In-domain and cross-domain argument position classification performance, reported as F-measure

Classifier	Gold	Automatic
PDTB-PDTB		
BoosTexter	98.12	97.81
PDTB-BioDRB		
BoosTexter	--	95.26

Source: Stepanov and Riccardi, 2014

2.2.4. Cross-Domain Argument Span Extraction

The in-domain performance of the argument span extraction models trained on PDTB sections 02-22 and tested on sections 23-24 is given on Table 1. The results are for 2 settings: 'Gold' and 'Auto'. In the 'Gold' settings the sentence splitting, tokenization and syntactic features are extracted from PTB, and in the 'Auto' they are extracted from automatic parse trees obtained using Stanford Parser (Klein and Manning, 2003).

In order to evaluate PDTB-BioDRB cross-domain performance we first evaluate the in-domain BioDRB argument span extraction. Since there is no gold sentence splitting, tokenization and syntactic parse trees, the models are trained using the features extracted from automatic parse trees. We use exactly the same feature sets as for PDTB models, which are optimized for PDTB. An important aspect is that in BioDRB the connective senses are different: there are 16 top level senses that are mapped to 4 top level PDTB senses. For the in-domain BioDRB models, the 16 senses were kept as is.

The results reported in Table 5 are average precision, recall and f-measure of 12-fold cross-validation. With respect to automatic sentence splitting, there are 717 inter-sentential and 1,919 intra-sentential relations (27% to 73%). Thus, BioDRB is less affected by PS *Arg1* performance than PDTB models, where the ratio is 619 to 976 (39% to 61%). Additionally, BioDRB PS *Arg1* performance is generally higher than that of PDTB. Overall, in-domain BioDRB argument extraction model performance is in-line with the PDTB models, with the exception that previous sentence *Arg2* has higher performance than the same sentence one.

Table 5: In-domain performance of the BioDRB-trained argument span extraction models

	Arg2			***Arg1***		
	P	**R**	**F1**	**P**	**R**	**F1**
AUTO						
SS	80.94	79.88	80.41	66.51	61.82	64.07
PS	82.99	82.99	82.99	57.50	55.62	56.53
ALL	81.45	80.67	81.06	63.87	60.00	61.87

Source: Stepanov and Riccardi, 2014

Both training and testing are on automatic sentence splitting, tokenization, and syntactic features. The results are reported for Same Sentence (SS) and Previous Sentence (PS)

models, and the joined results for each of the arguments (ALL) as average precision (P), recall (R), and F-measure (F1) of 12-fold cross-validation (Stepanov and Riccardi, 2014).

Similar to in-domain BioDRB argument span extraction, we perform 12 fold cross-validation for PDTB-BioDRB cross-domain argument span extraction. The cross-domain performance of the models described is given in the Table 6 under the 'Gold'. To make the cross-domain evaluation settings closer to the BioDRB in-domain evaluation, we additionally train PDTB models on the automatic features, i.e. features extracted from PDTB with automatic sentence splitting, tokenization and syntactic parsing.

Table 6: Cross-domain performance of the PDTB-trained argument span extraction models on Bio-DRB

	Arg2			***Arg1***		
	P	**R**	**F1**	**P**	**R**	**F1**
GOLD						
SS	80.37	76.58	78.42	60.82	56.40	58.52
PS	80.73	80.50	80.62	57.74	52.95	55.19
ALL	80.53	77.71	79.09	59.76	55.29	57.43
AUTO						
SS	77.60	75.05	76.30	60.76	55.21	57.83
PS	81.39	81.23	81.31	57.71	51.72	54.47
ALL	78.72	76.80	77.74	59.60	54.12	56.71

Source: Stepanov and Riccardi, 2014

For the 'Gold' setting the models from in-domain PDTB section are used. For 'Auto', the models are trained on automatic sentence splitting, tokenization, and syntactic features. The results are reported for Same Sentence (SS) and Previous Sentence (PS) models, and the joined results for each of the arguments (ALL) as average precision (P), recall (R), and F-measure (F1) of 12-fold cross-validation (Stepanov and Riccardi, 2014).

The first observation from cross-domain evaluation is that argument span extraction generalizes to biomedical domain much better that the discourse parsing subtasks of discourse connective detection and relation sense classification. Unlike those subtasks, the difference between in-domain BioDRB argument span extraction models and the models trained on PDTB is much less: e.g. for discourse connective detection the in-domain and cross-domain difference for BioDRB is 14 points (f-measures 69 and 55 in (Ramesh and Yu, 2010)), and for argument span extraction 2 and 4 points for Arg2 and Arg1 respectively. The difference between the models trained on automatic and gold parse trees is also not high, and gold feature trained models perform better with the exception of PS Arg2. Since training on automatic parse trees does not

improve cross-domain performance, the rest of the experiments are using gold features for training.

The two major differences between PDTB and BioDRB are vocabulary and connective senses. The out-of-vocabulary rate of PDTB on the whole BioDRB is 22.7% and of BioDRB on PDTB is 33.1%, which are very high. Thus, PDTB lexical features might not be very effective, and the models generalize well due to syntactic features. To test this hypothesis we train additional PDTB models on only syntactic features: POS-tags and IOB-chain and 'connective labels' – 'CONN' suffixed for the Beginning (B), Inside (I) or End (E) of the connective span, simulating discourse connective detection output.

Even though BioDRB connective senses can be mapped to PDTB, in (Prasad et al., 2011) it was observed that relation sense classification does not generalize well. To reduce the dependency of argument span extraction models on relation sense classification, the connective sense feature in the 'Baseline' models is also replaced by 'connective labels'. We train these models using gold features only, and, similar to previous experiments, do 12-fold cross-validation.

The performance of the adapted models is given in Table 7. The 'Syntactic' section gives the results of the models trained on syntactic features and the 'No Relation Sense' section gives the results for the models with 'connective labels' instead of connective senses, and the 'Baseline' repeats the performance of the PDTB-optimized models. The PDTB-optimized baseline, outperforms the adapted models on Arg2; however, 'No Relation Sense' Arg1 yields the best performance, and, though insignificantly, outperforms the baseline. Thus, the effect of replacing connective senses with 'connective labels' is negative for all cases except SS Arg1. Overall, the difference in performance between the 'Baseline' and 'No Relation Sense' models is an acceptable price to pay for the independence from relation sense classification. The most general models – unigrams of Part-of-Speech tags and IOB-chains together with 'connective labels' in the window of ±2 tokens – all have the performance lower than the baseline, which is expected given its feature set. However, for the easiest case of intra-sentential Arg2 it outperforms the model trained by replacing the connective sense in the baseline (i.e. 'No Relation Sense'). Degraded performance of Arg1 models indicates that lexical features are helpful. Introducing the tokens back into the 'Syntactic' model, and increasing the features to include also 2-grams, boosts the performance of the models to outperform the 'No Relation Sense' models in all but Previous Sentence Arg2 category. However, the models now yield performance comparable to the PDTB optimized baseline (insignificantly better), while being unaffected by poor cross-domain generalization of relation sense classification (see Table 8).

Table 7: Cross-domain performance of the PDTB-trained argument span extraction models on Bio-DRB

	Arg2			***Arg1***		
	P	**R**	**F1**	**P**	**R**	**F1**
Baseline						

SS	80.37	76.58	78.42	60.82	56.40	58.52
PS	80.73	80.50	80.62	57.74	52.95	55.19
ALL	80.53	77.71	79.09	59.76	55.29	57.43
Syntactic						
SS	82.00	75.03	78.33	61.07	51.80	56.01
PS	75.56	74.47	75.01	56.64	46.66	51.11
ALL	80.31	74.98	77.54	59.69	50.42	54.63
No Relation Sense						
SS	81.35	74.00	77.47	62.46	56.11	59.10
PS	80.35	80.13	80.24	57.58	52.25	54.74
ALL	81.16	75.67	78.30	60.86	54.87	57.69

Source: Stepanov and Riccardi, 2014

For the 'Syntactic' setting the models are trained on only syntactic features (POS-tag + IOB-chain) and 'connective labels'. For 'No Relation Sense', the models are trained by replacing connective sense with 'connective labels'. The 'Baseline' is repeated from Table 6. The results are reported for Same Sentence (SS) and Previous Sentence (PS) models, and the joined results for each of the arguments (ALL) as average precision (P), recall (R), and F-measure (F1) of 12-fold cross-validation (Stepanov and Riccardi, 2014).

Table 8: Cross-domain performance of the PDTB-trained argument span extraction model on unigram and bigrams of token, POS-tag, IOB-chain and 'connective label'

	Arg2			***Arg1***		
	P	**R**	**F1**	**P**	**R**	**F1**
SS	81.72	76.14	78.82	61.53	56.36	58.82
PS	80.31	79.84	80.07	58.55	52.82	55.44
ALL	81.27	77.10	79.12	60.56	55.30	57.80

Source: Stepanov and Riccardi, 2014

The results are reported for Same Sentence (SS) and Previous Sentence (PS) models, and the joined results for each of the arguments (ALL) as average precision (P), recall (R), and F-measure (F1) of 12-fold cross-validation (Stepanov and Riccardi, 2014).

The cross-domain argument extraction experiments indicate that models trained on PDTB-optimized feature set already have good generalization. However, they are dependent on relation sense classification task, which does not generalize well. By replacing connective

senses with ‘connective labels’ we obtain models independent of this task while maintaining comparable performance. The in-domain trained BioDRB models, however, perform better, as expected.

2.3. Analysis of Discourse Relations in Spoken Conversations

The two discourse relation annotated corpora we have worked so far are Penn Discourse Treebank (PDTB) and Biomedical Discourse Relation Bank. Both are essentially written monologues in English. Italian LUNA Corpus, on the other hand, contains discourse annotation on spoken dialogues in Italian. The issues of non-written-text and non-English discourse annotation were addressed in Tonelli et al. (2010). In this Section we address the issues of discourse parsing using spoken conversation corpus (LUNA).

2.3.1. Discourse in Speech and Text

The issues of discourse relation annotation of dialogs using Rhetorical Structure Theory (RST) Taboada and Mann (2006) are discussed in Stent (2000). The main difference between written text and a spoken dialog is in their segmentation into units of a discourse relations - connective and its arguments. While for written text there is only one speaker; thus, it is straightforward; the dialog introduces an additional level of segmentation - speakers and turns. As it was mentioned in Stent (2000), discourse relations may appear cross-speaker: different arguments of the same relation being in different speaker turns for elaboration relation, for instance. Additionally, due to the phenomena such as one speaker completing the other's utterance, even arguments may appear cross-speaker. Overall, in spoken dialogs the turn and speaker segmentation is not parallel to the discourse relation segmentation.

The developed PDTB-styled discourse parser essentially relies on the notions of sentence and adjacency. Dialogs, on the other hand, are segmented into turns. A turn may contain a part of a sentence or one or more sentences; and this information is generally not available. Turns, on the other hand, usually consist of one or several segments, partitioned with respect to some ‘event’ such as short silence, speech disfluency, or other. Taking any of these notions - turn or segment - as an equivalent for a sentence is equally problematic. We analyze the LUNA discourse annotation turn-wise to assess the ratio of discourse relations that are potentially processable by a discourse parser trained on text.

2.3.2. Data Analysis

Table 9 presents statistics on discourse relations in the LUNA Corpus. There are 1,052 explicit discourse relations in the LUNA Corpus (65.5% of total 1,606 annotated relations), which are signalled by 85 unique explicit discourse connectives. For comparison, in PDTB there are 18,459 explicit discourse relations and 100 unique explicit connectives. For PDTB and BioDRB we further analyzed discourse relations and connectives as inter- and intra-sentential. For LUNA Corpus, however, such analysis is not possible, since conversation transcriptions lack manual sentence segmentation, and there is no reference syntactic parses. However, unlike PDTB there are speaker and turn information. Since the discourse annotation procedure relied on the annotator's intuition for the disambiguation of overlapping turns and reconstruction of utterances, while speaker information is available in transcription layer of the corpus, we have

analyzed discourse relations as single-speaker and cross-speaker relations. A discourse relation consists of three spans: connective, Argument 1 and Argument 2; thus, the analysis additionally considers single vs. cross-speaker spans. Moreover, spontaneous dialogs contain interruptions; thus, some discourse relations may lack one or both of its arguments.

Table 9: Italian LUNA Corpus discourse annotation statistics

Annotation Statistic	Counts
Dialogues	60
Turns	3,750
Tokens	24,800
Total Relations	1,606
Explicit Relations	1,052
Unique Explicit Connectives	85
Unique Explicit Connective Surfaces	126

Source: partially from Tonelli et al., 2010

The statistics of Discourse Relation Span Analysis are given in Table 10. Since PDTB-styled discourse parser relies on the notion of sentence and essentially works mostly on intra-sentential discourse relations (since inter-sentential argument candidates are selected using heuristics), it is important to select a set of single-speaker single-turn relations. The ratio of such relations is only 37.6% (396), which is very low; thus, additional pre-processing for the reconstruction of discourse relations is required. For turn-wise analysis percent from single-speaker relations is given in parentheses.

Table 10: Italian LUNA Corpus discourse relation span statistics

	Counts	%
Total Explicit Relations	1,052	100.0%
Missing Span	19	1.8%
Speaker-wise Analysis		
Cross-Speaker Span	102	9.7%
Cross-Speaker Relation	170	16.2%
Single-Speaker Relation	763	72.5%
Turn-wise Analysis		
Multi-Turn Span	233	(30.5%) 22.2%
Single-Turn Span	530	(69.5%) 50.4%
Single-Turn Relation	396	(51.9%) 37.6%

2.3.3. Experiments and Results

The 60 human-human dialogs of LUNA Corpus that are annotated with discourse relation information are split into 3 sections. We use the first two sections for training and the third for testing. The distribution of data is such that 48 dialogs (794 relations) are used for training and 12 dialogs (258 relations) for testing.

The features used to train the LUNA discourse connective detection model are tokens (surface strings), part of speech tags and IOB-chains. The part-of-speech tags and IOB-chains are extracted from automatic syntactic parse trees using syntactic parser by Corazza et al. (2007).

For the experiments we considered a 'segment' to be equivalent to a sentence. The CRF model is trained taking these features in the +/-2 window.

Table 11: Discourse Connective Detection in LUNA Corpus.

Model	P	R	F1
LUNA: Token	64.87	28.88	40.91
LUNA: Token+POS+IOB	61.96	23.65	34.23
PDTB: Token (Pitler & Nenkova, 2009)	--	--	75.33
PDTB: Best (Pitler & Nenkova, 2009)	--	--	94.19
PDTB-PDTB (Ramesh & Yu, 2010)	88	81	84
BioDRB-BioDRB (Ramesh & Yu, 2010)	79	63	69
PDTB-BioDRB (Ramesh & Yu, 2010)	79	42	55

Results reported in terms of precision (P), recall (R) and F-measure (F1). PDTB and BioDRB in-domain and cross-domain results are given for a reference.

The discourse connective detection results are given in Table 11 for exact connective span match. For the other corpora we present published results by Pitler and Nenkova (2009) and Ramesh and Yu (2010). Their settings for PDTB are 10-fold cross-validation and for BioDRB 12-fold cross-validation. While Pitler and Nenkova (2009) makes use of complex syntactic features extracted from gold parse trees for the best classifier, Ramesh and Yu (2010) makes use of tokens, n-grams and morphological information only. The PDTB discourse connective model trained only on tokens (connective surfaces) already yields F-measure of 75.33. The LUNA model trained only on tokens yields the F-measure of 40.91. The interesting difference from the other corpora is that adding syntactic features results in a drop of performance of more than 6 points (from 40.91 to 34.23), which is indicative of the poor performance of the syntactic parser on the conversation data, as well as segments' being not appropriate for syntactic parsing. Thus, the syntactic parser should be adapted to speech data or the data should undergo the language style adaptation process. The next Section discussed the sentence boundary detection for conversation transcriptions.

2.4. Sentence Boundary Detection for Discourse Parsing of Spoken Conversations

Most discourse parsing approaches use paragraph, sentence or sub-sentence units as basic discourse units for analysis. However, unlike in text where paragraph segmentation is mostly trivial, sentence segmentation can be tackled with good success [Read et al. 2012; Gillick 2009] and subsentence units can be recovered with syntactic parsing, speech recordings come as a long stream of audio transcribed as a long stream of words, with the most frequent segmentation being speaker turns and based on long pauses. Sentence boundary detection consists in finding sentence ends in that stream of words. It is often related to punctuation prediction or dialog act tagging which aim at classifying the sentences between those boundaries. Both tasks can be tackled jointly.

State of the art approaches for sentence boundary detection and sentence classification mostly rely on hand-labeled, time-aligned transcripts of speech recordings, fed to a multiclass classifier. For the first task, each inter-word boundary is hypothesised as a sentence boundary

or non-boundary according to a feature vector extracted in the vicinity of the boundary. The second task can be tackled by extending the labelset of the first task (for instance “full-stop”, “question”, “comma”, “non-boundary”) [Favre et al., 2009] or extracting segment-level features after performing the first task [Quarteroni et al., 2011]. Relevant features used successfully in the past include lexical et pos-tag n-grams, pause duration, syntactic trees, and prosodic features such as F0 and energy contour, and phoneme and rhyme duration [Fung, 2011]. While pause duration and lexical features perform most of the job, prosody can add a small improvement; however the extraction of the latter is very affected by acoustic environment variability and speaker variability. Prosodic features are only effective when large amounts of training data are available.

In SENSEI, we study call-center recordings in order to improve conversation analysis in such context. Our study relies on the RATP-DECODA data which unfortunately does not include reliable sentence boundary marks, rendering all higher-level parsing processes unpractical because multiple sentences should not be processed together and said tools often imply a limit on the number of words they can process. In order to perform sentence boundary detection on the RATP-DECODA corpus, we turned to the TCOF corpus, made of recordings of conversations and interviews between French speakers about miscellaneous topics, and which has been carefully annotated for sentence-like units (full stops) and subsentence units (commas) [Wang et al, 2014]. This corpus contains about 250k word boundaries, of which 8% denote subsentence units and 6% denote sentence units.

We have trained 3-class CRFs in a fashion similar to that of [Liu et al. 2005] on the TCOF corpus. For features, we used pause-duration, extracted from time-aligning the corpus with jTrans tool from SYNALP, part-of-speech tags predicted with the macaon CRF tagger, trained on unpunctuated lowercased speech transcripts, and manually transcribed words. For part-of-speech and words, we rely on all n-grams of 1 to 3 items, with at least one item overlapping with the boundary. Performance is evaluated in term of recall, precision and F-score. The following table gives performances for the sentence-unit class on the TCOF test set, for each subgroup of features. It can be seen that pauses and words yield the best performance and that POS-tags bring a small improvement.

Table 12: Sentence-like unit segmentation performance on the TCOF corpus

TCOF test set	Recall	Precision	F-score
All features	61.11	57.81	59.41
Words only	42.42	48.46	45.24
POS-tags only	14.53	39.53	21.25
Pauses only	47.00	47.11	47.05

The next table summarizes the results on a subset of 100 conversations from RATP-DECODA which were manually annotated with sentence boundaries by two linguists who worked independently and then adjudicated their annotation.

Table 13: Sentence segmentation performance on the RATP-DECODA corpus

RATP-DECODA	Recall	Precision	F-score
Sentence-like units only	54.25	76.99	63.65
Both sentence-like units and subsentence units	57.51	72.64	64.19

Those results show that the task remains difficult but are comparable to results obtained in the meeting recording domain [Cuendet et al., 2007]. As future work, we plan on exploring the use of high-confidence boundaries to pre-segment the input before syntactic parsing and leave the task of final segmentation to the parser.

3. Extracting event and temporal structure from conversations

3.1 A combined approach to event detection

A major task within information extraction, event recognition has been successfully applied in research areas such as ontology generation, bioinformatics, news aggregation, business intelligence and text classification. Recognising events in these fields is generally carried out by means of pre-defined sets of relations, possibly structured into an ontology, which makes such tasks domain-dependent, but feasible.

Events can be expressed by various syntactic and lexical features of the text, such as the following:

- verbs and their arguments (e.g., "the committee rejected the proposal", "the country joined the EU");
- noun phrases headed by nominalizations (e.g., "economic growth", "Bulgaria's accession");
- adjective-noun combinations (e.g., "governmental measure");
- other event-referring nouns (e.g., "crisis").

We are providing an event detection component which we originally developed in the ARCOMEM FP7 project, are refining and adapting for SENSEI, and will improve through the course of the project. This component has been developed as a GATE [Cunningham et al., 2013] application and combines various approaches.

a) The top-down approach involves filling in templates relating to a range of events specified in advance; the tool uses relevant verbs (or other syntactic features mentioned above) to identify the events and their arguments to fill the slots. For example, an "election" event refers to a winning candidate or party and may also contain a date and location. This approach generally gives high precision but low recall.

b) The bottom-up approach identifies verbal relations in the text and classifies them into semantic categories; new events can be inferred from these. This approach generally gives higher recall but lower precision.

Both approaches are based on linguistic preprocessing (tokenization, sentence splitting, POS tagging, morphological analysis, and verb and noun phrase chunking) followed by named entity recognition.

The top-down recognition subcomponent uses a semantic approach to finding the verbal expressions which represent the relations. We automatically create lists of verbs for each relation, using information from WordNet and VerbNet to group verbs into semantic categories. The tool then applies rules to find the relevant verbs and their arguments in the text. As mentioned above, this approach produces high precision -- but it also needs to contain the right verbs for the texts and domain.

The bottom-up subcomponent is also rule-based, but the rules are flexible and underspecified, based on syntactic structure and semantic relations from WordNet [Fellbaum, 1998] rather than predefined relations. This part uses noun phrase and verb phrase chunking, identifies linguistic patterns around verb phrases, and clusters the verbs into semantically related categories to find new relations.

This bottom-up technique involving open-domain IE can find previously unknown events and is not limited to a predefined set of relations; this allows the discovery of new information and compliments the top-down approach. It has been tested with interesting results on news texts in ARCOMEM.

The combined approach described here is being integrated into the SENSEI prototype; after evaluation on SENSEI data it will be retuned appropriately for the relevant domains. Formal evaluations using standard methods (precision and recall) will be carried out as necessary (and reported) as part of the research during SENSEI to improve and extend this component.

3.2 Recognising TimeML Events and Times

TimeML [Pustejovsky et al., 2003] is an XML language for temporally annotating events and temporal expressions in natural language. It defines events very broadly as situations that happen or occur, or elements describing states or circumstances in which something obtains or holds the truth [Llorens et al., 2010]; events are usually (but not always) expressed by verbs and nominalizations.

The TimeML guidelines define 7 classes of event:

- reporting: action of a person or organisation declaring or narrating an event (e.g. "say");
- perception: physical perception of another event (e.g. "see", "hear");
- aspectual: aspectual predication of another event (e.g. "start", "continue");
- I_Action: intensional action (e.g. "try");
- I_State: intensional state (e.g. "feel", "hope");
- state: circumstance in which something holds the truth (e.g. "war", "in danger");
- occurrence: events that describe things that happen (e.g. "erupt", "arrive").

USFD has previously carried out some research (in the ARCOMEM project) on identifying TimeML events, principally occurrences and intensional actions. In that project we investigated the EVITA system [Sauri et al., 2005], which is freely available and has a good record on the TimeML eventrecognition task. Its performance was good in our experiments but it would have been very difficult to integrate into the ARCOMEM prototype for technical reasons. EVITA may be worth reconsidering in SENSEI because the prototype environment is more amenable to integrating a Python-based component and the documents in the conversational repository will be stored with POS tags and related information that EVITA requires.

We will also experiment with state of the art tools for timex (temporal expression) recognition, including the following:

- The HeidelTime rule-based system has good precision but not very high recall. USFD is planning to collaborate with its main developer in 2015 on adapting it for social media text (https://code.google.com/p/heideltime/);
- The TIPSem system uses the CRF (Conditional Random Fields) machine learning technique [Llorens et al., 2010] and has a good reputation for event as well as timex recognition. It has been applied successfully to English and Spanish;
- The TIMEN rule-based system carries out timex normalization, i.e., generating normalized TIMEX3 annotations (in TimeML) from temporal expressions in natural language (http://code.google.com/p/timen/).

4. Intra-document coreference for conversations & social media

Task 4.3 is concerned with the development of intra-document coreference algorithms using the BART platform and tuned for conversational and social media data. In this first year work on the task involved the revision or creation of datasets for the task for English and Italian, and work on domain adaptation of the state of the art algorithms in BART –trained on news—for social media. We discuss each aspect of the work in turn.

4.1 Corpus checking and annotation

Two datasets of anaphora in social media were used in this first year. For Italian, the existing Live Memories Anaphora Blog corpus was used. This corpus had been previously annotated (Rodriguez et al., 2010) but not used on a large scale, so it was checked. For English, a new dataset was created annotating part of the data collected for the Social Media Summarization Shared Task at MULTILING.

The annotation scheme used for the social media datasets (English and Italian) is a variant of the LiveMemories annotation scheme (Rodriguez et al., 2010) which in turn is based on the ARRAU annotation scheme (Poesio and Artstein, 2008). In this corpus all noun phrases are treated as mentions, and the entire noun phrase is considered. All anaphoric relations of identity between any mentions are annotated. Coordinations are also treated as mentions, and annotated.

4.1.1 Correcting the Live Memories Anaphora Blogs

The Live Memories Blogs corpus (LM-Blogs) consists of three sets of documents in Italian: (a) Jurka/Samba, (b) Uncommented Blogs/Trento Blogs and (c) Commented Blogs/Click, Eco, Polis, Queer Blogs, for a total of 95 documents containing 71 392 words (see table 2). All three sets had to be re-annotated due to inconsistencies found in previous annotations. The annotation was carried out by an experienced annotator native speaker of Italian, following the annotation scheme described above and it was carried out in three stages, one for each set of documents. For the annotation itself, the MMAX21 (Müller and Strube, 2006) annotation tool was used which defines a stand-off XML format particularly well-suited for representing rich linguistic annotations.

[1] http://mmax2.sourceforge.net/

Table 14: LM Blogs corpus: document set names and sizes.

Set of Document	Number of Documents	Number of Words
Jurka / Samba	30	22 266
Uncommented Blogs / Trento Blogs	35	18 075
Commented Blogs / Click, Eco, Polis, Queer Blogs	30	31 051

In order to ensure the quality of the annotations, the same methodology developed within the EVALITA initiative2 was employed, which consists of performing several iterations of automatic consistency checks via scripts followed by revisions from the annotator.

4.1.2 Intra-doc coreference annotation of the SENSEI Social Media Corpus

The SENSEI Social Media Corpus (SSMC) will be a subset of the data for the Social Media Summarisation (SMS) Task including freely distributable news articles and readers' comments from The Guardian for English, the blog of Vittorio Zucconi (hosted in La Repubblica) and/or Corriere della Sera for Italian and Le Monde for French. The size of the SMS dataset will be approximately 75K words per language.

The Social Media Summarisation (SMS) Task is organised as a pilot track within an established shared task on Multilingual Multi-document Summarisation, MultiLing 2015. The purpose of the SMS task is to set the ground for investigating how the large volume of readers' comments as appearing nowadays on most on-line news publishers (e.g., The Guardian) can be summarised. The SMS task is described in detail in Deliverable D7.1., here we briefly discuss how the data was selected.

There are a number of desirable features that were sought when selecting the data for the SMS task. Firstly, we needed news articles rich in readers' comments as well as topic variety (e.g., politics, environment, sports, etc.). Then, we needed to be able to sift comments according to various parameters such as number of likes, number of replies, chronologically and comment length. Finally, there are two approaches for selecting the data top-down by browsing through URLs by hand and using news aggregators such as the NewsExplorer (Steinberger et al., 2011), or bottom-up by extracting data scraped automatically from news sources. In both cases, data is stored in XML files produced by Websays.

The XML format of Websays presented in detail in Deliverable D2.1., is a conveniently flat sequence of 'clippings' where every clipping represents either a post (i.e., news article) or a single reader's comment, crafted for the purposes of processing large volumes of data. Data such as news article title, text, associated comments (preserving comment hierarchy, i.e., replies of comments), number of likes per comment and metadata such as, posting date, author, source, etc., are scraped from on-line news sources and dumped into XML files. Then these XML files are further processed, so that data can be extracted based on the desirable

[2] http://www.evalita.it

features described above (e.g., select the 10 news articles with greatest number of comments and choose the top 50 comments according to number of likes). At this stage data is stored into UTF-8-encoded text files preserving original layout (i.e., comments indentation).

Once the data set is selected and saved in UTF-8-encoded text files, then these are processed with BART's pipeline (explained below) to get linguistic annotation in MMAX2 format. After this is achieved the same procedure as for the LM-Blog annotation, explained in the previous section, is employed.

Table 15: SENSEI Social Media corpus: document set names and sizes

Set of Document	Number of Documents	Number of Words
Data from The Guardian (English)	10	35 490
Data from Repubblica (Italian)	*to be collected in due course	-
Data from Le Monde	*to be collected in due course	-

4.2 Adapting intra-document coreference to social media

4.2.1 Domain Adaptation with BART

As a starting point we took the latest version of the Beautiful Anaphora Resolution Toolkit3 (BART), version 2.0, and a version for Italian which implements the Unstructured Information Management Architecture (UIMA), and was developed as part of the project LiMoSINe4. The key advantages of using BART is that it provides a generic framework for machine learning experiments on coreference, and hence, it is particularly well-suited for incorporating new models, new languages, and domain adaptation experiments.

We have identified three core milestones in our approach to domain adaptation:

a) Baseline performance: computing baseline results by training coreference models on standard newswire datasets for English and Italian and testing them on the new social media domain without any adaptation
b) Ceiling performance: computing upper bound performance by a standard ten-fold cross validation on the old domain (newswire) only (the newswire domain is an easier domain for coreference than the social media domain, which is a more challenging domain for most higher-level NLP tasks)
c) New domain performance: we have subdivided this in two; one training and testing new models on the new domain only (limited labelled data, hence, constricted performance) and two, taking advantage of all data available new and old domain to train models and testing on the new domain only

[3] http://www.bart-anaphora.eu
[4] http://limosine-project.eu

Progress on each of the aforementioned milestones is described below.

4.2.2 Baseline results for Social Media

Table 4.3 shows baseline results for models trained on the old domain (the train set of the Evalita corpus) and tested on the new domain, the LM-Blog corpus. As can be seen from the table, the baseline results are quite low oscillating around an F1 of 40%. This alone shows that the social media domain is harder than the newswire domain, but also that training on the old domain alone is not sufficient to achieve a coreference performance that is usable in higher level tasks such as summarisation, where usually an F1 of 60+ is required to bring in tangible benefits.

Table 16: Baseline results for Italian

Set of Document	Recall (%)	Precision (%)	F-1 (%)
Jurka / Samba	43.3	40.4	41.8
Uncommented Blogs / Trento Blogs	44.1	39.2	41.6
Commented Blogs / Click, Eco, Polis, Queer Blogs	42.2	41.1	41.7

4.2.3 Ceiling results for Domain Adaptation

Upper performance bounds will be produced by running BART on a data set representing the old domain, a standard newswire dataset. The underlying assumption is that, since the new social media domain is a harder and less studied domain than the newswire domain, a good estimate of performance on the old domain is a good upper bound for performance on the new domain. The best method for estimating performance is by running several train/test iterations and computing the average. An instance of such train/test cycle is a 10-fold cross validation where the test sets are mutually exclusive, hence, upper bound performance will be produced by 10-X-Validation on the Evalita train data (another subset of the data produced by the Live Memories project).

4.2.4 Experiments with Domain Adaptation

Our next steps are to test models in the new domain trained on both data from the new domain as well as data from the old domain, thus taking advantage of all the data available. The key research questions to be addressed are what is the best way to perform domain adaptation with minimal effort for the task of coreference and how to make the most of the data available (which is usually very limited for new unexplored domains).

5. CONCLUSIONS

In this deliverable we presented the work on discourse analysis carried out during the first year of the project. Three main strands of work were pursued: discourse parsing of spoken

conversations in French and Italian, event extraction from English texts and intra-document coreference in social media in Italian and English.

On discourse parsing, the pipeline developed by Stepanov & Riccardi was tested for cross-domain and genre generalization. Discourse parsing is broken down into discourse relation detection, argument position classification, argument span extraction, and relation sense classification. For the sub-tasks of discourse relation detection and relation sense classification the attained performance is high (F1 ~ 90+%), whereas for the other two, inherently harder tasks, reported performances ranged in the 80s. Then, various domain adaptation experiments for all sub-tasks were carried out showing that domain adaptation for the problem at hand is challenging, but feasible and that further research is needed to attain performance equivalent to that in the original domain. In the context of spoken conversations, an additional problem was explored, that of sentence boundary detection for discourse parsing where initial experimental results were presented showing that the task is difficult and suggesting as next steps to use high-confidence boundaries to pre-segment the input before syntactic parsing and leave the task of final segmentation to the parser.

In the work on temporal and event structure, an event detection component developed in ARCOMEM was adapted to the new domain, combining two approaches: a top-down, template filling approach, and a bottom-up, verbal relations-driven approach searching for the best trade-off between precision and recall. More experiments will be carried out with state-of-the-art tools for timex (temporal expression) recognition.

In the work on intra-document coreference, an existing resource (the Blog subcorpus of the LiveMemories Anaphora corpus of intra-document coreference in Italian) was used to adapt the latest, state-of-the-art version of the BART toolkit on social media data. A new dataset for English was also created, collecting data from the Guardian via the SENSEI tools developed by WebSays and annotating them using the same guidelines. The next steps are to carry out extensive machine learning experiments on domain adaptation for coreference, taking advantage of both small data sets from the new domain and large collections from the old domain which are readily available (in particular, for English). This has the potential of breaking new ground in research on coreference, as both domain and language adaptation in this area are largely underexplored.

REFERENCES

[Artiles et al., 2007] Artiles, J., Gonzalo, J., and Sekine, S. (2007) "The SemEval-2007 WePS evaluation: Establishing a benchmark for the web people search task", In Proc. of SemEval.

[Artstein and Poesio, 2008] Artstein, R. and Poesio, M. (2008) "Intercoder agreement for Computational Linguistics", Computational Linguistics, 34(4).

[Asher and Lascarides, 2003] Asher, N. and Lascarides, A. (2003) The Logic of Conversation. Cambridge University Press.

[Bagga and Baldwin, 1998] Bagga, A., Baldwin, B. (1998) "Entity-based cross-document coreferencing using the vector space model", In Proc. of COLING/ACL.

[Baker et al., 1998] Baker, F.C., Fillmore, J.C., and Lowe, B.J. (1998) "The Berkeley FrameNet project", In Proc. of COLING/ACL.

[Barzilay and Lee, 2004] Barzilay, R. and Lee, L. (2004) "Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization", In Proc. of NAACL-HLT.

[Bazillon et al., 2012] Bazillon, T., Deplano, M., Bechet, F., Nasr, A., and Favre, B. (2012) "Syntactic annotation of spontaneous speech: application to call-center conversation data", In Proc. of LREC.

[Bechet and Nasr, 2009] Bechet, F. and Nasr, A. (2009) "Robust dependency parsing for Spoken Language Understanding of spontaneous speech", In Proc. of INTERSPEECH.

[Bechet et al., 2012] Bechet, F., Maza, B., Bigouroux, N., Bazillon, T., El-Bèze, M., De Mori, R., and Arbillot, E. (2012) "DECODA: a call-center human-human spoken conversation corpus", In Proc. of LREC.

[Benzitoun et al., 2012] Benzitoun, C., Fort, K., and Sagot, B. (2012). TCOF-POS: un corpus libre de français parlé annoté en morphosyntaxe. In Actes de la conf´erence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 99–112.

[Blitzer et al., 2006] Blitzer, J., McDonald, R., and Pereira, F. (2006) "Domain Adaptation with Structural Correspondence Learning", In Proc. of EMNLP.

[Byrd et al., 2008] Byrd, R.J., Neff, M.S., Teiken, W., Park, Y., Cheng, K.S.F, Gates, S.C., and Visweswariah, K. (2008) "Semi-automated logging of contact center telephone calls", In Proc. of CIKM.

[Carlson et al., 2001] Carlson, L., Marcu, D., and Okurowski, M.E. (2003) "Building a discourse-tagged corpus in the framework of rhetorical structure theory". In J. Kuppevelt and R. Smith (eds) Current Directions in Discourse and Dialogue. Kluwer.

[Chambers and Jurafsky, 2011] Chambers, N. and Jurafsky, D. (2011) "Template-based information extraction without the templates", In Proc. of ACL.

[Chen and Martin, 2007] Chen, Y. and Martin, J. (2007) "Towards robust unsupervised personal name disambiguation", In Proc. of EMNLP.

[Coppola et al., 2009] Coppola, B., Moschitti, A., and Riccardi, G. (2009) "Shallow Semantic Parsing for Spoken Language Understanding", In Proc. of NAACL.

[Corazza et al., 2007] Anna Corazza, Alberto Lavelli, and Giorgio Satta. Analisi sintatticastatistica basata su costituenti. Intelligenza Artificiale, 4(2):38-39, 2007.

[Csomai and Mihalcea, 2008] Csomai, A. and R. Mihalcea (2008) "Linking documents to encyclopedic knowledge", IEEE Intelligent Systems.

[Cuendet et al., 2007] Cross-Genre Feature Comparisons for Spoken Sentence Segmentation, Sebastien Cuendet, Dilek Hakkani-Tür, Elizabeth Shriberg, James Fung, Benoit Favre, International Journal of Semantic Computing (IJSC), Volume 1, Issue 3, 2007.

[Cunningham et al., 2013] H. Cunningham, V. Tablan, A. Roberts, K. Bontcheva (2013) Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLoS Comput Biol 9(2): e1002854. doi:10.1371/journal.pcbi.1002854

[Daumé, 2007] Daumé III H. (2007) "Frustratingly Easy Domain Adaptation", In Proc. of ACL.

[Dinarelli et al., 2009] Dinarelli, M., Quarteroni, S., Tonelli, S., Moschitti, A., and Riccardi, G. (2009) "Annotating Spoken Dialogs: from Speech Segments to Dialog Acts and Frame Semantics", In Proc. of EACL Workshop on Semantic Representation of Spoken Language.

[Faiz and Mercer, 2013] Syeed Ibn Faiz and Robert E Mercer. Identifying explicit discourse connectives in text. In Advances in Artificial Intelligence, pages 64-76. Springer, 2013.

[Favre et al., 2009] Benoit Favre, Dilek Hakkani-Tür, Elizabeth Shriberg, "Syntactically-informed Models for Comma Prediction", IEEE International Conference in Acoustics, Speech and Signal Processing (ICASSP), Taipei (Taiwan), 2009.

[Fellbaum, 1998] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[Fung, 2011], Automatic Design of Prosodic Features for Sentence Segmentation, James Fung, 2011, Thesis, UC Berkeley.

[Gillick 2009] Sentence boundary detection and the problem with the US, Dan Gillick, HLT 2009.

[Ghosh et al., 2011] Sucheta Ghosh, Richard Johansson, Giuseppe Riccardi, and Sara Tonelli. Shallow discourse parsing with conditional random fields. In Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011), 2011.

[Klein and Manning, 2003] Dan Klein and Christopher D. Manning. Fast exact inference with a factored model for natural language parsing. In Advances in Neural Information Processing Systems 15 (NIPS 2002), pages 3-10, 2003.

[Lafferty et al., 2001] John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of 18th International Conference on Machine Learning, pages 282{289, 2001.

[Lin et al., 2012] Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. A pdtb-styled end-to-end discourse parser. Natural Language Engineering, 1:1-35, 2012.

[Llorens et al., 2010] Llorens, H., Saquete, H., and Navarro-Colorado, B. TimeML events recognition and classification: learning CRF models with semantic roles. s.l. : Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational

Linguistics. 2010.

[Llorens et al., 2010] Llorens, Hector, Estela Saquete, and Borja Navarro. "Tipsem (English

and Spanish): Evaluating crfs and semantic roles in tempeval-2." Proceedings of the 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics, 2010.

[Müller and Strube, 2006] Christoph Müller, Michael Strube (2006): Multi-Level Annotation of Linguistic Data with MMAX2. In: Sabine Braun, Kurt Kohn, Joybrato Mukherjee (Eds.): Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods. Frankfurt: Peter Lang, pp. 197-214. (English Corpus Linguistics, Vol.3).

[Pitler and Nenkova, 2009] Emily Pitler and Ani Nenkova. Using syntax to disambiguate explicit discourse connectives in text. In Proceedings of the ACL-IJCNLP Conference, 2009.

[Poesio and Artstein 2008] Poesio, M. and Artstein, R. 2008. Anaphoric annotation in the ARRAU corpus. Proc. Of LREC.

[Prasad et al., 2008] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The penn discourse treebank 2.0. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), 2008.

[Prasad et al., 2011] Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. The biomedical discourse relation bank. BMC Bioinformatics, 12(1):188, 2011.

[Pustejovsky et al., 2003] TimeML: Robust specification of event and temporal expressions in text. Pustejovsky, J., Castano, J., Ingria, R., Sauri, R., Gaizauskas, R., Setzer, A., Katz, G., Radev, D. s.l. : New directions in question answering, 3, 28-34., 2003.

[Quarteroni et al., 2011] Simultaneous dialog act segmentation and classification from human-human spoken conversations, Quarteroni, Silvia and Ivanov, Alexei V and Riccardi, Giuseppe, Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, 2011.

[Ramesh et al., 2012] Balaji Polepalli Ramesh, Rashmi Prasad, Tim Miller, Brian Harrington, and Hong Yu. Automatic discourse connective detection in biomedical text. Journal of the American Medical Informatics Association, 19(5):800-808, 2012.

[Rodriguez et al 2010] Rodriguez, K., F. Delogu, Y. Versley, E. W. Stemle, and M. Poesio, 2010. Anaphoric annotation of Wikipedia and Blogs in the Live Memories Corpus. Proc. Of LREC 2010.

[Read et al. 2012] Sentence Boundary Detection: A Long Solved Problem? J. Read, R. Dridan, S. Oepen, LJ. Solberg, COLING 2012.

[Sauri et al., 2005] Roser Saurí, Robert Knippen, Marc Verhagen and James Pustejovsky. 2005. Evita: A Robust Event Recognizer for QA Systems. Short Paper. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP) 2005: pages 700-707.

[Steinberger et al., 2011] Steinberger Ralf, Sylvia Ombuya, Mijail Kabadjov, Bruno Pouliquen, Leonida Della Rocca, Jenya Belyaeva, Monica De Paola & Erik van der Goot (2011). Expanding a multilingual media monitoring and information extraction tool to a new language: Swahili. Language Resources and Evaluation Journal, Volume 45, Issue 3, pp. 311-330 (DOI 10.1007/s10579-011-9165-9).

[Wang et al, 2014] Macrosyntactic Segmenters of a French spoken corpus, Wang, Ilaine and Kahane, Sylvain and Tellier, Isabelle, LREC 2014.