Abductive Reasoning with a Large Knowledge Base for Discourse Processing

Ekaterina Ovchinnikova
University of Osnabrück
eovchinn@uos.de

Niloofar Montazeri
USC ISI
niloofar@isi.edu

Theodore Alexandrov
University of Bremen
theodore@uni-bremen.de

Jerry R. Hobbs
USC ISI
hobbs@isi.edu

Michael C. McCord
IBM Research
mcmccord@us.ibm.com

Rutu Mulkar-Mehta
USC ISI
me@rutumulkar.com

Abstract

This paper presents a discourse processing framework based on weighted abduction. We elaborate on ideas described in Hobbs et al. (1993) and implement the abductive inference procedure in a system called Mini-TACITUS. Particular attention is paid to constructing a large and reliable knowledge base for supporting inferences. For this purpose we exploit such lexical-semantic resources as WordNet and FrameNet. We test the proposed procedure and the obtained knowledge base on the Recognizing Textual Entailment task using the data sets from the RTE-2 challenge for evaluation. In addition, we provide an evaluation of the semantic role labeling produced by the system taking the Frame-Annotated Corpus for Textual Entailment as a gold standard.

1 Introduction

In this paper, we elaborate on a semantic processing framework based on a mode of inference called *abduction*, or inference to the best explanation. In logics, abduction is a kind of inference which arrives at an explanatory hypothesis given an observation. Hobbs et al. (1993) describe how abductive reasoning can be applied to the discourse processing problem viewing the process of interpreting sentences in discourse as the process of providing the best explanation of why the sentence would be true. In this framework, interpreting a sentence means 1) proving its logical form, 2) merging redundancies where possible and 3) making assumptions where necessary. As the reader will see later in this paper, abductive reasoning as a discourse processing technique helps to solve many pragmatic problems such as reference resolution, the interpretation of noun compounds, the resolution of some kinds of syntactic and semantic ambiguity as a by-product. We adopt this approach. Specifically, we use a system we have built called *Mini-TACITUS*[1] (Mulkar et al., 2007) that provides the expressivity of logical inference but also allows probabilistic, fuzzy, or defeasible inference and includes measures of the "goodness" of abductive proofs and hence of interpretations of texts and other situations.

The success of a discourse processing system based on inferences heavily depends on a knowledge base. The main contribution of this paper is in showing how a large and reliable knowledge base can be obtained by exploiting existing lexical semantic resources and can be successfully applied to reasoning tasks on a large scale. In particular, we experiment with axioms extracted from WordNet, see Fellbaum (1998), and FrameNet, see Ruppenhofer et al. (2006). In axiomatizing FrameNet we rely on the study described in Ovchinnikova et al. (2010).

We evaluate our inference system and the obtained knowledge base in recognizing textual entailment (RTE). As the reader will see in the following sections, inferences carried out by Mini-TACITUS are fairly general and not tuned for a particular application. We decided to test our approach on RTE because this is a well-defined task that captures major semantic inference needs across many natural language processing applications, such as question answering, information retrieval, information extraction, and document summarization. For evaluation, we have chosen the RTE-2 data set (Bar-Haim et al., 2006), because besides providing text-hypothesis pairs and a gold

[1] *http://www.rutumulkar.com/download/TACITUS/tacitus.php*

standard this data set has been annotated with FrameNet frames and role labels (Burchardt and Pennacchiotti, 2008) which gives us the possibility of evaluating our frames and role labeling based on the axioms extracted from FrameNet.

This paper is structured as follows: In section 2 we briefly introduce our natural language pipeline, explain how abductive reasoning can be applied to discourse processing, and say a few words about the Mini-TACITUS system. In section 3 we describe the obtained knowledge base. Section 4 presents our procedure for recognizing textual entailment. In section 5 we provide an evaluation of our reasoning pipeline on the RTE-2 data set. The last section concludes the paper and gives an outlook on future work and perspectives.

2 NL Pipeline and Abductive Reasoning

Our natural language pipeline produces interpretations of texts given the appropriate knowledge base. A text is first input to the English Slot Grammar (ESG) parser (McCord (1990, 2010)). For each segment, the parse produced by ESG is a dependency tree that shows both surface and deep structure. The deep structure is exhibited via a word sense predication for each node, with logical arguments. These logical predications form a good start on a logical form (LF) for the whole segment. An add-on to ESG converts the parse tree into a LF in the style of Hobbs (1985). The LF is a conjunction of predications, which have generalized entity arguments that can be used for showing relationships among the predications. These LFs are used by the downstream components.

The interpretation of the text is carried out by an inference system called Mini-TACITUS using weighted abduction as described in detail in Hobbs et al. (1993). Mini-TACITUS tries to prove the logical form of the text, allowing assumptions where necessary. Where the system is able to prove parts of the logical form, it is anchoring it in what is already known from the overall discourse or from a knowledge base. Where assumptions are necessary, it is gaining new information. Obviously, there are many possible proofs in this procedure. A cost function on proofs enables the system to chose the "best" (the cheapest) interpretation. The key factors involved in assigning a cost are the following:

1. Proofs with fewer assumptions are favored.
2. Short proofs are favored over long ones.
3. Salient and plausible axioms are favored over less salient or less plausible axioms.
4. Proofs are favored that exploit the inherent implicit redundancy in texts.

Let us illustrate the procedure with a simple example. Suppose that we want to construct the best interpretation of the sentence *John composed a sonata.* As a by-product, the procedure will disambiguate between two readings of *compose*, namely between the "form" reading instantiated for example in the sentence *Three representatives composed a committee*, and the "create art" meaning instantiated in the given sentence. After being processed by the parser, the sentence will be assigned the following logical form where the numbers (20) after every proposition correspond to the default costs of these propositions.[2] The total cost of this logical form is equal to 60.

John(x1):20 & *compose*(e1,x1,x2):20 & *sonata*(x2):20

Suppose our knowledge base contains the following axioms.

1) *form*(e0,x1,x2):90 → *compose*(e0,x1,x2)
2) *create_art*(e0,x1,x2):50 & *art_piece*(x2):40 → *compose*(e0,x1,x2)
3) *art_piece*(x1):90 → *sonata*(x1)

Unlike deductive axioms, abductive axioms should be read "right to left". Thus, the propositions on the right hand side (*compose*, *sonata*) correspond to an input, whereas the left hand side propositions will be assumed given the input. The number assigned to each proposition on the left hand side shows what percentage of the total input cost the assumption of this proposition will cost.[3] For example, if the proposition *compose* costs 20 then the assumption of *form* will cost 18.

[2] The actual value of the default costs of the input propositions does not matter, because, as the reader will see in this section, the axiom weights which affect the costs of the resulting interpretations are given as *percentages* from the input proposition costs. The only heuristic we use here concerns setting all costs of the input propositions to be equal (all propositions cost 20 in the discussed example). This heuristic needs a further investigation to be approved or modified.

[3] The axiom weights in the given example are arbitrary.

Two interpretations can be constructed for the given logical form. The first one is the result of the application of axioms 1 and 3. Note that the costs of the backchained propositions (*compose*, *sonata*) are set to 0, because their costs are now carried by the newly introduces assumptions (*form*, *art_piece*). The total cost of the first interpretation **I1** is equal to 56.

I1: *John*(x1):20 & *compose*(e1,x1,x2):0 & *sonata*(x2):0 & *form*(e1,x1,x2):18 & *art_piece*(x2):18

The second interpretation is constructed in two steps. First, axioms 2 and 3 are applied as follows.

I2$_1$: *John*(x1):20 & *compose*(e1,x1,x2):0 & *sonata*(x2):0 &
create_art(e1,x1,x2):10 & *art_piece*(x2):8 & *art_piece*(x2):18

The total cost of **I2**$_1$ is equal to 56. This interpretation is redundant, because it contains the proposition *art_piece* twice. The procedure will merge propositions with the same predicate, setting the corresponding arguments of these propositions to be equal and assigning the minimum of the costs to the result of merging. The idea behind such mergings is that if an assumption has already been made then there is no need to make it again. The final form of the second interpretation **I2**$_2$ with the cost of 38 is as follows. The "create art" meaning of *compose* has been brought forward because of the implicit redundancy in the sentence which facilitated the disambiguation.

I2$_2$: *John*(x1):20 & *compose*(e1,x1,x2):0 & *sonata*(x2):0 & *create_art*(e1,x1,x2):10 &
art_piece(x2):8

Thus, on each reasoning step the procedure 1) applies axioms to propositions with non-zero costs and 2) merges propositions with the same predicate, assigning the lowest cost to the result of merging. Reasoning terminates when no more axioms can be applied.[4] The procedure favors the cheapest interpretations. Among them, the shortest proofs are favored, in the sense that if two interpretations have the same cost then the one which has been constructed with fewer axiom application steps is considered to be "better".

It is easy to see that changing weights of axioms can crucially influence the reasoning process. Axiom weights can help to propagate more frequent and reliable inferences as well as to distinguish between "real" abduction and deduction. For example, an axiom backchaining from *dog* to *animal* should in the general case have a weight below 100, because it is cheap to assume that there is an animal if there is a dog; it is a reliable deduction. On the contrary, assuming *dog* given *animal* should have a weight above 100.

In order to avoid undesirable mergings, we introduce non-merge constraints. For example, in the sentence *John reads a book and Bill reads a book* the two *read* propositions should not be merged because they refer to different actions. This is ensured by the following non-merge constraint: if not all arguments of two propositions (which are not nouns) with the same predicate can be merged, then these propositions cannot be merged. The constraint implies that in the sentence above two *read* propositions cannot be merged, because *John* being the first argument of the first *read* cannot be merged with *Bill*.[5] This constraint is a heuristic; it corresponds to the intuition that it is unlikely that the same noun refers to different objects in a short discourse, while for other parts of speech it is possible. An additional corpus study is needed in order to prove or disprove it.

The described procedure provides solutions to a whole range of natural language pragmatics problems. The best explanation for a vague predicate, such as those conveyed by prepositions, verbs like "have", and the implicit relation between the nouns in a nominal compound, is generally some more specific predicate derivable from the surrounding text or the knowledge base, and this solves the problem of discovering the specific meanings of general words in context. Similarly, discovering discourse structure is a matter of coming up with a specific relation, such as causality or similarity, conveyed by the adjacency of two segments of discourse. Moreover, this account of interpretation solves the problem of where to stop drawing inferences, which could easily be unlimited in number; an inference is appropriate if it is part of the lowest-cost proof of the logical form.

Adapting Mini-TACITUS to Large-Scale Knowledge Base

Mini-TACITUS (Mulkar et al., 2007) began as a simple backchaining theorem-prover intended to be a more transparent version of the original TACITUS system, which was based on Stickel's PTTP

[4] In practice, we use the depth parameter d and do not allow an inference chain with more that d steps.

[5] Recall that only propositions with the same predicate can be merged, therefore *John* and *Bill* cannot be merged.

system (Stickel, 1988). Originally, Mini-TACITUS was not designed for treating large amounts of data. A clear and clean reasoning procedure rather than efficiency was in the focus of its developers. In order to make the system work with the large-scale knowledge base, we had to perform several optimization steps and add a couple of new features.

For avoiding the reasoning complexity problem, we have introduced two parameters. The time parameter t is used to restrict the processing time. After the processing time exceeds t the reasoning terminates and the best interpretation so far is output. The time parameter ensures that an interpretation will be always returned by the procedure even if reasoning could not be completed in a reasonable time. The depth parameter d restricts the depth of the inference chain. Suppose that a proposition p occurring in the input has been backchained and a proposition p' has been introduced as a result. Then, p' will be backchained and so on. The number of such iterations cannot exceed d. The depth parameter reduces the number of reasoning steps.

Since Mini-TACITUS processing time increases exponentially with the input size (sentence length and number of axioms), making such a large set of axioms work was an additional issue. For speeding up reasoning it was necessary to reduce both the number of the input propositions and the number of axioms. In order to reduce the number of axioms, a two-step reduction of the axiom set is performed. First, only the axioms which could be evoked by the input propositions or as a result of backchaining from the input are selected for each reasoning task. Second, the axioms which could never lead to any merging are filtered out. Concerning the input propositions, those which could never be merged with the others (even after backchaining) are excluded from the reasoning process.

3 Knowledge Base

As described in the previous section, the Mini-TACITUS inferences are based on a knowledge base (KB) consisting of a set of axioms. In order to obtain a reliable KB with a sufficient coverage we have exploited existing lexical-semantic resources.

First, we have extracted axioms from WordNet (Fellbaum, 1998), version 3.0, which has already proved itself to be useful in knowledge-intensive NLP applications. The central entity in WordNet is called a *synset*. A synset is a set of cognitive synonyms belonging to the same part of speech (nouns, verbs, adjectives or adverbs). Synsets correspond to word senses, so that every lexeme can participate in several synsets. For every word sense, WordNet indicates the frequency of this particular word sense in the WordNet annotated corpora. We have used the lexeme-synset mapping for generating axioms, with the corresponding frequencies of word senses converted into the axiom weights. For example, in the axioms below, the verb *compose* is mapped to its sense 2 in WordNet which participates in *synset-X*.

compose-2(e1,x1,x2):80 → *compose*(e1,x1,x2)
synset-X(e0,e1):100 → *compose-2*(e1,x1,x2)

Moreover, we have converted the following WordNet relations defined on synsets into axioms: hypernymy, instantiation, entailment, similarity, meronymy. Hypernymy and instantiation relations presuppose that the related synsets refer to the same entity (the first axiom below), whereas other types of relations relate synsets referring to different entities (the second axiom below). All axioms based on WordNet relations have the weights equal to 100.

synset-1(e0,e1):100 → *synset-2*(e0,e1)
synset-1(e0,e1):100 → *synset-2*(e2,e3)

WordNet also provides morphosemantic relations defined on word senses, for example *buy-buyer*. These relations relate verbs and nouns. WordNet distinguishes between 14 types of such relations.[6] We use these relation types in order to define the direction of the entailment and map the arguments in a proper way. For example, the "agent" relation (*buy-buyer*) stands for a bi-directional entailment such that the noun is the first (agentive) argument of the verb:

buy-1(e0,x1,x2):100 → *buyer-1*(x1)
buyer-1(x1):100 → *buy-1*(e0,x1,x2)

Additionally, we have exploited the WordNet synset definitions. In WordNet the definitions are

[6] *http://wordnet.princeton.edu/wordnet/download/standoff/*

given in natural language form. We have used the extended WordNet resource[7] which provides logical forms for the definition in WordNet version 2.0. We have adapted logical forms from extended WordNet to our representation format and converted them into axioms; for example the following axiom represents the meaning of the synset containing such lexemes as *horseback*. These axioms have the total weight of 100.

on(e2,e1,x2):25 & *back*(e3,x2):25 & *of*(e4,x2,x1):25 & *horse*(e5,x1):25 → *synset-X*(e0,x0)

The second resource which we have used as a source of axioms is FrameNet, release 1.5, see Ruppenhofer et al. (2006). FrameNet has a shorter history in NLP applications than WordNet, but lately more and more researchers have been demonstrating its potential to improve the quality of question answering (Shen and Lapata, 2007) and recognizing textual entailment (Burchardt et al., 2009). The lexical meaning of predicates in FrameNet is represented in terms of frames which describe prototypical situations spoken about in natural language. Every frame contains a set of roles corresponding to the participants of the described situation. Predicates with similar semantics are assigned to the same frame; e.g. both *give* and *hand over* refer to the GIVING frame. For most of the lexical elements FrameNet provides syntactic patterns showing the surface realization of these lexical elements and their arguments. Syntactic patterns also contain information about their frequency in the FrameNet annotated corpora. We have used the patterns and the frequencies for deriving axioms such as for example the following.

GIVING(e1,x1,x2,x3):70 & DONOR(e1,x1):0 & RECIPIENT(e1,x2):0 & THEME(e1,x3):0 →
give(e1,x1,x3) & *to*(e2,e1,x2)
HIRING(e1,x1,x3):90 & EMPLOYER(e1,x1) & EMPLOYEE(e1,x3) →
give(e1,x1,x2,x3):10 & *job*(x2)

The first pattern above corresponds to the phrases like *John gave a book to Mary* and the second – less frequent – to phrases like *John gave Mary a job*. It is interesting to note that application of such axioms provides a solution to the problem of semantic role labeling as a by-product. As in the statistical approaches, more frequent patterns will be favored. Moreover, patterns helping to detect implicit redundancy will be brought forward.

FrameNet also introduces semantic relations defined on frames such as inheritance, causation or precedence; for example the GIVING and GETTING frames are connected with the causation relation. Roles of the connected frames are also linked, e.g. DONOR in GIVING is linked with SOURCE in GETTING. Frame relations have no formal semantics in FrameNet. In order to generate corresponding axioms, we have used the previous work on axiomatizing frame relations and extracting new relations from corpora (Ovchinnikova et al., 2010). Weights of the axioms derived from frame relations depend on corpus-based similarity of the lexical items assigned to the corresponding frames. An example of an axiomatized relation is given below.[8]

GIVING(e0,x1,x2,x3):120 & DONOR(e0,x1):0 & RECIPIENT(e0,x2):0 & THEME(e0,x3):0 &
causes(e0,e1):0 →
GETTING(e1,x2,x3,x1) & SOURCE(e1,x1) & RECIPIENT(e1,x2) & THEME(e1,x3)

Both WordNet and FrameNet are manually created resources which ensures a relatively high quality of the resulting axioms as well as the possibility of exploiting the linguistic information provided for structuring the axioms. Although manual creation of resources is a very time-consuming task, WordNet and FrameNet, being long-term projects, have an extensive coverage of English vocabulary. The coverage of WordNet is currently larger than that of FrameNet (155 000 vs. 12 000 lexemes). However, the fact that FrameNet introduces complex argument structures (roles) for frames and provides mappings of these structures makes FrameNet especially valuable for reasoning.

The complete list of axioms we have extracted from these resources is given in table 1.

[7] *http://xwn.hlt.utdallas.edu/*

[8] The "causes" predicate is supposed to be linked to an underlying causation theory, see for example *http://www.isi.edu/∼hobbs/bgt-cause.text*. However, in the described experimental settings we have left the abstract theories out and evaluated only the axioms extracted from the lexical-semantic resources.

Table 1: Statistics of extracted axioms

Axiom type	Source	Numb. of axioms
Lexeme-synset mappings	WN 3.0	422,000
Lexeme-synset mappings	WN 2.0	406,000
Synset relations	WN 3.0	141,000
Derivational relations	WN 3.0 (annotated)	35,000
Synset definitions	WN 2.0 (parsed, annotated)	120,500
Lexeme-frame mappings	FN 1.5	50,000
Frame relations	FN 1.5 + corpora	6,000

4 Recognizing Textual Entailment

As the reader can see from the previous sections, the discourse processing procedure we have presented is fairly general and not tuned for any particular type of inferences. We have evaluated the procedure and the KB derived from WordNet and FrameNet on the Recognizing Textual Entailment (RTE) task, which is a generic task that seems to capture major semantic inference needs across many natural language processing applications. In this task, the system is given a text and a hypothesis and must decide whether the hypothesis is entailed by the text plus commonsense knowledge.

Our approach is to interpret both the text and the hypothesis using Mini-TACITUS, and then see whether adding information derived from the text to the knowledge base will reduce the cost of the best abductive proof of the hypothesis as compared to using the original knowledge base only. If the cost reduction exceeds a threshold determined from a training set, then we predict entailment.

A simple example would be the text *John gave a book to Mary* and the hypothesis *Mary got a book*. Our pipeline constructs the following logical forms for these two sentences.

T: *John*(x1):20 & *give*(e1,x1,x2):20 & *book*(x3):20 & *to*(e2,e1,x3):20 & *Mary*(x3):20
H: *Mary*(x1):20 & *get*(e1,x1,x2):20 & *book*(x2):20

These logical forms constitute the Mini-TACITUS input. Mini-TACITUS applies the axioms from the knowledge base to the input logical forms in order to reduce the overall cost of the interpretations. Suppose that we have three FrameNet axioms in our knowledge base. The first one maps *give to* to the GIVING frame, the second one maps *get* to GETTING and the third one relates GIVING and GETTING with the causation relation. The first two axioms have the weights of 90 and the third 120. As a result of the application of the axioms the following best interpretations will be constructed for T and H.

I(T): *John*(x1):20 & *give*(e1,x1,x2):0 & *book*(x3):20 & *to*(e2,e1,x3):0 & *Mary*(x3):20 &
GIVING(e0,x1,x2,x3):18
I(H): *Mary*(x1):20 & *get*(e1,x1,x2):0 & *book*(x2):20 & GETTING(e0,x1,x2):18

The total cost of the best interpretation for H is equal to 58. Now the best interpretation of T will be added to H with the zero costs (as if T has been totally proven) and we will try to prove H once again. First of all, merging of the propositions with the same names will result in reducing costs of the propositions *Mary* and *book* to 0, because they occur in T:

I(T+H): *John*(x1):0 & *give*(e1,x1,x2):0 & *book*(x3):0 & *to*(e2,e1,x3):0 & *Mary*(x3):0 &
GIVING(e0,x1,x2,x3):0 & *get*(e1,x1,x2):0 & GETTING(e0,x1,x2):18

The only proposition left to be proved is GETTING. Using the GETTING-GIVING relation as described in the previous section, this proposition can be backchained on to GIVING which will merge with GIVING coming from the T sentence. H appears to be proven completely with respect to T; the total cost of its best interpretation given T is equal to 0. Thus, using knowledge from T helped to reduce the cost of the best interpretation of H from 58 to 0.

The approach presented does not have any special account for logical connectors such as *if*, *not*, *or* etc. Given a text *If A then B* and a hypothesis *A and B* our procedure will most likely predict entailment. Thus, at the moment our RTE procedure mainly accounts for the informational content of the text fragments, being able to detect the "aboutness" overlap of T and H. In our framework, a fuller treatment of the logical structure of the natural language would presuppose a more complicated strategy of merging redundancies.

5 Evaluation Results

We have evaluated our procedure on the RTE-2 dataset [9], see Bar-Haim et al. (2006) . The RTE-2 dataset contains the development and the test set, both including 800 text-hypothesis pairs. Each dataset consists of four subsets, which correspond to typical success and failure settings in different applications: information extraction (IE), information retrieval (IR), question answering (QA), and summarization (SUM). In total, 200 pairs were collected for each application in each dataset.

As a baseline we have processed the datasets with an empty knowledge base. Then we have done 2 runs, first, using axioms extracted from WordNet 3.0 plus FrameNet, and, second, using axioms extracted from the WordNet 2.0 definitions. In both runs the depth parameter was set to 3. The development set was used to train the threshold as described in the previous section.[10]

Table 2 contains results of our experiments.[11] Accuracy was calculated as the percentage of pairs correctly judged. The results suggest that the proposed method seems to be promising as compared to the other systems evaluated on the same task. Our best run gives 63% accuracy. Two systems participating the RTE-2 Challenge had 73% and 75% accuracy, two systems achieved 62% and 63%, while most of the systems achieved 55%-61%, cf. Bar-Haim et al. (2006).

For our best run (WN 3.0 + FN), we present the accuracy data for each application separately (table 2). The distribution of the performance of Mini-TACITUS on the four datasets corresponds to the average performance of systems participating in RTE-2 as reported by Garoufi (2007). The most challenging task in RTE-2 appeared to be IE. QA and IR follow, and finally, SUM was titled the "easiest" task, with a performance significantly higher than that of any other task.[12]

It is worth noting that the performance of Mini-TACITUS increases with the increasing time of processing. This is not surprising. We use the time parameter t for restricting the processing time. The smaller t is, the fewer chances Mini-TACITUS has for applying all relevant axioms. The experiments carried out suggest that optimizing the system computationally could lead to producing significantly better results. Tracing the reasoning process, we found out that given a long sentence and a short processing time Mini-TACITUS had time to construct only a few interpretations, and the real best interpretation was not always among them.

The lower performance of the system using the KB based on axioms extracted from extended WordNet can be also easily explained. At the moment we define non-merge constraints (see section 2) for the input propositions only. The axioms extracted from the synset definitions introduce a lot of new lexemes into the logical form, since these axioms define words with the help of other words rather than abstract concepts. These new lexemes, especially those which are frequent in English, result in undesired mergings (e.g., mergings of frequent prepositions), since no non-merge constraints are defined for them. In order to fix this problem, we will need to implement dynamic non-merge constraints which will be added on the fly if a new lexeme is introduced during reasoning. The WN 3.0 + FN axiom set does not fall into this problem, because these axioms operate on frames and synsets rather than on lexemes.

In addition, for the run using axioms derived from FrameNet, we have evaluated how well we do in assigning frames and frame roles. For Mini-TACITUS, semantic role labeling is a by-product of constructing the best interpretation. But since this task is considered to be important as such in the NLP community, we provide an additional evaluation for it. As a gold standard we have used the Frame-Annotated Corpus for Textual Entailment, FATE, see Burchardt and Pennacchiotti (2008). This corpus provides frame and semantic role label annotations for the RTE-2 challenge test set.[13]

[9] *http://pascallin.ecs.soton.ac.uk/Challenges/RTE2/*

[10] Interpretation costs were normalized to the number of propositions in the input.

[11] "Time" stands for the value of the time parameter – processing time per sentences, in minutes; "Numb. of ax." stands for average number of axioms per sentence.

[12] In order to get a better understanding of which parts of our KB are useful for computing entailment and for which types of entailment, in future, we are planning to use the detailed annotation of the RTE-2 dataset describing the source of the entailment which was produced by Garoufi (2007). We would like to thank one of our reviewers for giving us this idea.

[13] FATE was annotated with the FrameNet 1.3 labels, while we have been using 1.5 version for extracting axioms. However, in the new FN version the number of frames and roles increases and there is no message about removed frames in the General Release Notes R1.5, see *http://framenet.icsi.berkeley.edu*. Therefore we suppose that most of the frames and roles used for the FATE annotation are still present in FN 1.5.

Table 2: Evaluation results for the RTE-2 test set

KB	Accuracy	Time	Numb. of ax.	
			T	H
No KB	57%	1	0	0
WN 3.0 + FN	62%	20	533	237
WN 3.0 + FN	63%	30	533	237
Ext. WN 2.0	60%	20	3700	1720
Ext. WN 2.0	61%	30	3700	1720

Task	Accuracy
SUM	75%
IR	64%
QA	62%
IE	50%

Table 3: Evaluation of frames/roles labeling towards FATE

System	Frame match	Role match	
	Recall	Precision	Recall
Shalmaneser	0.55	0.54	0.37
Shalmaneser + Detour	0.85	0.52	0.36
Mini-TACITUS	0.65	0.55	0.30

It is important to note that FATE annotates only those frames which are relevant for computing entailment. Since Mini-TACITUS makes all possible frame assignments for a sentence, we provide only the recall measure for the frame match and leave the precision out.

The FATE corpus was also used as a gold standard for evaluating the Shalmaneser system (Erk and Pado, 2006) which is a state-of-the-art system for assigning FrameNet frames and roles. In table 2 we replicate results for Shalmaneser alone and Shalmaneser boosted with the WordNet Detour to FrameNet (Burchardt et al., 2005). The WN-FN Detour extended the frame labels assigned by Shalmaneser with the labels related via the FrameNet hierarchy or by the WordNet inheritance relation, cf. Burchardt et al. (2009). In frame matching, the number of frame labels in the gold standard annotation that can also be found in the system annotation (recall) was counted. Role matching was evaluated only on the frames that are correctly annotated by the system. The number of role labels in the gold standard annotation that can also be found in the system annotation (recall) as well as the number of role labels found by the system which also occur in the gold standard (precision) were counted.[14] Table 3 shows that given FrameNet axioms, the performance of Mini-TACITUS on semantic role labeling is compatible with those of the system specially designed to solve this task.

6 Conclusion and Future Work

This paper presents a discourse processing framework underlying the abductive reasoner called *Mini-TACITUS*. We have shown that interpreting texts using weighted abduction helps solve pragmatic problems in discourse processing as a by-product. In this paper, particular attention was paid to the construction of a large and reliable knowledge base populated with axioms extracted from such lexical-semantic resources as WordNet and FrameNet. The reasoning procedure as well as the knowledge base were evaluated in the Recognizing Textual Entailment task. The data for evaluation were taken from the RTE-2 Challenge. First, we have evaluated the accuracy of the entailment prediction. Second, we have evaluated frame and role labeling using the Frame-Annotated Corpora for Textual Entailment as the gold standard. In both tasks our system showed performance compatible with those of the state-of-the art systems. Since the inference procedure and the axiom set are general and not tuned for a particular task, we consider the results of our experiments to be promising concerning possible manifold applications of Mini-TACITUS.

The experiments we have carried out have shown that there is still a lot of space for improving the procedure. First, for successful application of Mini-TACITUS on a large scale the system needs to be

[14] We do not compare filler matching, because the FATE syntactic annotation follows different standards as the one produced by the ESG parser, which makes aligning fillers non-trivial.

computationally optimized. In its current state, Mini-TACITUS requires too much time for producing satisfactory results. We use the time parameter t for controlling the processing time. If reasoning takes longer than t, Mini-TACITUS is required to terminate and output the best interpretation so far. Thus, the time limit can prevent the system from producing the real best interpretation. As our experiments suggest (cf. table 2), speeding up reasoning may lead to significant improvements in the system performance. Since Mini-TACITUS was not originally designed for large-scale processing, its implementation is in many aspects not effective enough. We hope to improve it by changing the data structure and re-implementing some of the main algorithms.

Second, in the future we plan to elaborate our treatment of natural language expressions standing for logical connectors such as implication *if*, negation *not*, disjunction *or* and others. Quantifiers such as *all*, *each*, *some* also require a special treatment. This advance is needed in order to achieve more precise entailment inferences, which are at the moment based in our approach on the core information content ("aboutness") of texts. Concerning the heuristic non-merge constraints preventing undesired mergings as well as the heuristic for assigning default costs (see section 2), in the future we would like to perform a corpus study for evaluating and possibly changing these heuristics.

Another future direction concerns the enlargement of the knowledge base. Hand-crafted lexical-semantic resources such as WordNet and FrameNet provide both an extensive lexical coverage and a high-value semantic labeling. However, such resources still lack certain features essential for capturing some of the knowledge required for linguistic inferences. First of all, manually created resources are in principle static; updating them with new information is a slow and time-consuming process. By contrast, commonsense knowledge and the lexicon are dynamic entities which undergo daily updates. New words appear and existing words acquire new senses and associations. For accommodating all this dynamic knowledge, static knowledge bases seem to be not the best solution. In our reasoning procedure, we plan to make use of the distributional similarities of words in a large Web-corpus such as for example Wikipedia. The idea of word similarity spaces comes back to the distributional hypothesis claiming that words occurring in the same contexts tend to have similar meanings (Harris, 1954). Many researchers working on RTE have already been using word similarity for computing similarity between texts and hypotheses, see for example Mehdad et al. (2010). In our approach, we plan to incorporate word similarities into the reasoning procedure making them affect proposition costs so that propositions implied by the context (similar to other words in the context) will become cheaper to prove. This extension might give us a performance improvement in RTE, because it will help to relate those propositions from H for which there are no appropriate axioms in the knowledge base to propositions in T.

Lexical-semantic resources and word similarity spaces as knowledge sources for reasoning share a common shortcoming: They imply too little structure. WordNet and FrameNet enable some argument mappings of related synsets or frames, whereas word spaces operate on bare words. FrameNet seems to introduce more semantic structure than any of the other existing lexical resources, but even FN relations are not always enough for drawing relevant inferences, cf. Ovchinnikova et al. (2010). For example, for resolving the T-H pair *X suffers from a disease* - *X has a disease* a complex axiomatization of the predicates *have* and *suffer* in the "medical" context is required. We are engaged in two types of efforts to obtain more structured knowledge. The first effort is the manual encoding of abstract theories explicating concepts that pervade natural language discourse, such as causality, change of state, and scales, and the manual encoding of axioms linking lexical items to these theories. A selection of the core theories can be found at *http://www.isi.edu/ hobbs/csk.html*. The second effort concerns making use of the existing ontologies. In computer science, an ontology is a formal representation of a domain of knowledge as a set of concepts and relationships between these concepts (Gruber, 2009). The main purpose of such ontologies is to enable reasoning over their content, i.e. to enable answering queries about concepts, relations and their instances. The recent progress of the Semantic Web technologies has stimulated extensive development of the domain-specific ontologies represented in Web Ontology Language (OWL) as well as development of inference machines specially designed to reason with OWL representations.[15] In practice, domain-specific ontologies usually represent detailed and structured knowledge about particular domains (e.g. geography, medicine etc.). We intend to make Mini-TACITUS able to use this knowledge through querying an externally stored ontology with the help of an existing OWL-reasoner. This extension will give us a possibility to access elaborated domain-specific knowledge which might be crucial for interpretation of domain-specific texts.

We believe that implementation of the mentioned improvements and extensions will make Mini-

[15] *www.w3.org/2001/sw/*, *www.w3.org/TR/owl-features/*, *http://www.cs.man.ac.uk/ sattler/reasoners.html*.

TACITUS a powerful reasoning system equipped with enough knowledge to solve manifold NLP tasks on a large scale. In our view, the experiments with the axioms extracted from the lexical-semantic resources presented in this paper show the potential of weighted abduction for natural language reasoning and open new ways for its application.

References

Bar-Haim, R., I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, and I. Szpektor (2006). The second PASCAL recognising textual entailment challenge. In *Proc. of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Burchardt, A., K. Erk, and A. Frank (2005). A WordNet Detour to FrameNet. In *Sprachtechnologie, mobile Kommunikation und linguistische Resourcen*, Volume 8.

Burchardt, A. and M. Pennacchiotti (2008). FATE: a FrameNet-Annotated Corpus for Textual Entailment. In *Proc. of LREC'08*.

Burchardt, A., M. Pennacchiotti, S. Thater, and M. Pinkal (2009). Assessing the impact of frame semantics on textual entailment. *Natural Language Engineering 15*(4), 527–550.

Erk, K. and S. Pado (2006). Shalmaneser - a flexible toolbox for semantic role assignment. In *Proc. of LREC'06*, Genoa, Italy.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database* (First ed.). MIT Press.

Garoufi, K. (2007). Towards a better understanding of applied textual entailment: Annotation and evaluation of the rte-2 dataset. Master's thesis, Saarland University.

Gruber, T. (2009). Ontology. In *Encyclopedia of Database Systems*, pp. 1963–1965.

Harris, Z. (1954). Distributional structure. *Word 10*(23), 146–162.

Hobbs, J. R. (1985). Ontological promiscuity. In *Proceedings, 23rd Annual Meeting of the Association for Computational Linguistics*, Chicago, Illinois, pp. 61–69.

Hobbs, J. R., M. Stickel, and P. Martin (1993). Interpretation as abduction. *Artificial Intelligence 63*, 69–142.

McCord, M. C. (1990). Slot grammar: A system for simpler construction of practical natural language grammars. In *Natural Language and Logic: International Scientific Symposium, Lecture Notes in Computer Science*, pp. 118–145. Springer Verlag.

McCord, M. C. (2010). Using Slot Grammar. Technical report, IBM T. J. Watson Research Center. RC 23978Revised.

Mehdad, Y., A. Moschitti, and F. M. Zanzotto (2010). Syntactic/semantic structures for textual entailment recognition. In *Proc. of HLT '10: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 1020–1028.

Mulkar, R., J. R. Hobbs, and E. Hovy (2007). Learning from Reading Syntactically Complex Biology Texts. In *Proc.of the 8th International Symposium on Logical Formalizations of Commonsense Reasoning. Palo Alto*.

Ovchinnikova, E., L. Vieu, A. Oltramari, S. Borgo, and T. Alexandrov (2010). Data-Driven and Ontological Analysis of FrameNet for Natural Language Reasoning. In *Proc. of LREC'10*, Valletta, Malta.

Ruppenhofer, J., M. Ellsworth, M. Petruck, C. Johnson, and J. Scheffczyk (2006). FrameNet II: Extended Theory and Practice. *International Computer Science Institute*.

Shen, D. and M. Lapata (2007). Using Semantic Roles to Improve Question Answering. In *Proc. of EMNLP-CoNLL*, pp. 12–21.

Stickel, M. E. (1988). A prolog technology theorem prover: Implementation by an extended prolog compiler. *Journal of Automated Reasoning 4*(4), 353–380.