Towards a Better Understanding of Applied Textual Entailment

Annotation and Evaluation of the RTE-2 Dataset

Konstantina Garoufi

A thesis presented for the degree of
Master of Science in Language Science and Technology
at the Saarland University

UNIVERSITÄT
DES
SAARLANDES

September 2007

Author: Konstantina Garoufi
Dept. of Computational Linguistics and Phonetics
Saarland University
Saarbrücken, Germany
dgarf@coli.uni-sb.de

Advisors: Prof. Dr. Manfred Pinkal
Dr. Stefan Thater

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Saarbrücken, 27. September 2007

Konstantina Garoufi

Abstract

Applied textual entailment is a newly introduced generic empirical task that captures major semantic inferences across a wide spectrum of Natural Language Processing applications. In the present thesis we quest for a better understanding of the task by means of investigating a benchmark dataset for textual entailment, the dataset of the Second PASCAL Recognising Textual Entailment (RTE-2) Challenge.

We propose a scheme for annotation of textual entailment, the Annotating RTE (ARTE) scheme, which models a range of diverse entailment mechanisms. The annotation of a considerable portion of the RTE-2 dataset following this scheme enables us an evaluation of the textual entailment data by gaining insights into the semantic-linguistic properties of the textual entailment phenomenon.

Based on this evaluation, the thesis finally examines from various aspects the performance of the textual entailment systems participating in the RTE-2 Challenge, relative to different types of entailment. The methodology followed and the resulting observations make first steps towards a thorough analysis of systems' performance, which is a key issue for the advancement of textual entailment technology.

Acknowledgements

This work would probably not have been possible—and definitely not so much fun for me—without the care and consideration of several people, whom I cannot thank enough.

To my advisor Prof. Dr. Manfred Pinkal I owe first of all the idea for this study, as well as the means to realize it. Beyond that, I owe him the gratification of experiencing such moments of insight and inspiration, as only a man of remarkable sharpness of mind and rare perceptiveness can instigate. He is a model of advisor.

I am no less indebted to my other advisor, Dr. Stefan Thater, for his enormous motivation, encouragement and kindness. With his scripts, often on-the-fly and always efficient, he showed me the real magic of a computational linguist in action; with his constant support and optimism he taught me faith, confidence and joy in my work.

The infinite patience and warmheartedness of Olga Kukina have been instrumental in allowing the annotation project to be completed. I am thankful for that, and for all those endless, stimulating linguistic explorations she shared with me.

It would be difficult to overstate my thankfulness to Rui Wang, who contributed to my work in so many substantial ways. Our eternal discussions about textual entailment, research, and both, have hardly ever been anything less than thrilling, and have never ceased to be one of my sources of enthusiasm.

A series of very interesting conversations with Aljoscha Burchardt and Dr. Marco Pennacchiotti has significantly broadened my views on the general topic of this thesis. Moreover, I thank them warmly for taking the time to read my draft pages, and return them to me filled with useful comments and suggestions.

Many thanks also to Dr. Sebastian Padó and Prof. Dr. Anette Frank for kindly offering valuable advice, as well as posing challenging questions and making constructive criticism, during some of the thesis' explorative quests.

I cannot but be full of gratitude to all my friends, whose support, reaching me undiminished from ranging geographical distances, I have been so lucky as to abundantly enjoy through it all. Very special thanks are due to Kateryna Ignatova, who magically managed to help me find energy, strength and courage even at the hardest times.

Finally, quite cliché as it may sound, I am always deeply grateful for my family's everlasting love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Most readers would probably agree that the truth of the hypothesis H can be inferred from the truth of the text T in the example below.

T *If destruction of the rainforest continues, global warming will continue to take place.*

H *Destruction of the rainforest affects the world's weather.*

— 295, PASCAL RTE-2 test set

The ability to draw such textual inferences is a fundamental component of human cognition, whose success leans on the human aptitude for handling the variability of language. If Natural Language Processing (NLP) applications are to meet real-life demands, they naturally also need to learn how to model this ability.

This important task, called textual entailment recognition, is the main object of the recently started PASCAL Recognizing Textual Entailment Challenge, an initiative to promote a generic evaluation framework for real-world textual entailment systems. The investigation of the textual entailment datasets released in the frame of this initiative is the purpose of the present thesis, ultimately aiming at a better understanding of the textual entailment phenomenon.

This chapter introduces the task of textual entailment recognition and illustrates the motivation for the thesis, as well as its main contributions.

1.1 Textual Entailment

According to a classical definition of entailment in formal semantics, as in (Chierchia and McConnell-Ginet, 2000),

> *a text T entails a hypothesis H, if H is true in every circumstance (possible world) in which T is true.*

This kind of definition, however, imposes a strictness that is rather inappropriate for many practical situations regarding NLP, in which uncertain but highly plausible inferences are still useful.

The problem is addressed by the notion of **applied textual entailment**[1], as defined by Dagan and Glickman (2004), which takes an operational approach based on empirical evaluation. By this definition,

> *a text T entails a hypothesis H, if, typically, a human reading T would infer that H is most likely true.*

The advantages of such a perspective for NLP are straightforward: the evaluation is performed using a human gold standard, as in other NLP tasks, and at the same time, common background knowledge is assumed, in the way also expected from applications. Therefore the notion is purposefully restricted to such an informal definition, so as to match the equally uncertain nature of the task.

Only a short list of concrete textual entailment annotation guidelines, introduced by Glickman (2006), complement the above definition:

- *Entailment is a directional relation; the hypothesis must be entailed by the text, but the opposite is not required.*
- *The hypothesis must be fully entailed by the text and not include parts that cannot be inferred.*
- *Cases in which inference is very probable (but not absolutely certain) should be judged as* true.
- *Common background knowledge that is typical for a reader of the given type of texts is presupposed; on the other hand, the presumption of highly specific knowledge is unacceptable.*

[1] Note that the use of the term *textual entailment* in this general sense has been criticized (e.g. Zaenen et al. (2005), Manning (2006)). The wider notion of *local textual inference* was proposed as more appropriate; nonetheless the former seems fairly well-established in the research community.

1.2 The PASCAL Recognizing Textual Entailment Challenge

Drawing from the idea of applied textual entailment, the PASCAL Network of Excellence recently started an attempt to promote a generic evaluation framework covering semantic-oriented inferences needed for practical applications.

The launch of the Recognizing Textual Entailment (RTE) Challenge (Dagan et al., 2006) aimed at setting a benchmark for the development and evaluation of semantic methods that typically address the same types of problems but in different, application-oriented manners. As many of the needs of several applications can be cast in terms of textual entailment, the ultimate goal is to promote the development of general entailment recognition engines, designed to provide generic modules across applications.

The initiative has been widely embraced, yielding to this day three successful yearly RTE challenges—the first (RTE-1; 2005), second (RTE-2; 2006) and third (RTE-3; 2007)[2]—, as well as an increasingly high interest in the research community. In this frame, which has developed a character more explorative rather than competitive, participating systems are required to judge the entailment value of short pairs of text snippets (text T and hypothesis H), like the one presented in the beginning of this chapter. The notion of entailment considered for this purpose is explicitly the one of applied textual entailment, as defined in Section 1.1.

Submissions have been numerous and diverse, evaluated for accuracy (the percentage of correctly judged pairs) and, optionally, average precision, as a measure for the ranking of pairs according to their entailment confidence (when applicable). The methods typically employed by the participating systems include similarity measures between T and H, cross-pair similarity measures, detection of mismatch, and, to a limited extent, logical inference.

The systems' results demonstrate significant general improvement with time, with overall accuracy levels ranging from 50% to 60% on RTE-1 (17 submissions), from 53% to 75% on RTE-2 (23 submissions), and from 49% to 80% on RTE-3 (26 submissions).

1.3 The RTE Datasets

Clearly, what plays a most central role on this applied account of textual entailment is the nature of the datasets involved, with data collection and

[2] http://www.pascal-network.org/Challenges/{RTE, RTE2, RTE3}.

annotation processes largely determining important parameters like the complexity of the task and its correspondence to real-life application settings.

For this reason, the datasets provided by the RTE Challenge organizers are intended to include typical T–H pairs that correspond to success and failure cases of actual text processing applications, dealing with tasks such as Information Extraction (IE), Information Retrieval (IR), Question Answering (QA) and Summarization (SUM). They are divided into two balanced corpora: the development (dev) set, released early so as to provide typical examples of the task requirements, and the test set, released a few weeks prior to systems' submission. The systems' results are evaluated exclusively on the test set.

The collected pairs are intended to challenge systems on how they handle a broad range of textual entailment phenomena. To achieve that, they strive for representing a range of different levels of entailment reasoning, including syntactic, lexical, logical reasoning and world knowledge, at different degrees of difficulty. The specific approaches, however, taken with respect to the compilation of the datasets, lead to certain observations.

In practice, the datasets are manually compiled by human annotators, using existing application-specific system resources or the output of Web-based systems, with a focus on the general domain of news. On this process the annotators are instructed to obtain a reasonable balance among the different types of pairs, but no concrete indications regarding the type each pair may correspond to are available, as the pairs are constructed on the fly.

For instance, as the organizers of RTE-1 Dagan et al. (2006) observe, the annotators' selection policy evidently yielded more negative entailment pairs than positive ones in the cases where T and H had a very high degree of lexical overlap in RTE-1, resulting from their bias to avoid high correlation between word overlap and entailment. Dagan et al. add that they are not in a position to provide any information about the distribution of different entailment factors in the RTE-1 datasets, or make any direct predictions about the performance of participating systems in particular applications.

On an attempt to quantitatively determine the presence of mere paraphrases in the same datasets, Bayer et al. (2005) report that 94% of the dev2 set of RTE-1 consists of paraphrases, as opposed to classic entailments. They remark that it seems unclear how RTE annotation techniques could possibly be applied to corpora for creating a good balance of different representative types of inferences, or what should generally be understood as such a balance.

On top of that, simple lexical overlap reportedly (Bar-Haim et al., 2006) achieved an accuracy of as high as 60% on RTE-2, interestingly outperforming some more sophisticated lexical methods. MacCartney et al. (2006) notice that the inferences involved in the data are, from a human perspective, fairly superficial, as no long chains of reasoning are present, and higher-level reasoning arises only occasionally.

Finally, despite the increased maturity of the task gained from experience, the RTE-3 organizers Giampiccolo et al. (2007) highlight the urging need for theoretical refinements in order to overcome current limitations. In particular, as they point out, the arbitrary distributions of the pairs has come to constitute a major problem for the advancement of the field, which needs to be addressed by both refining and extending the data generation and evaluation methodologies currently applied.

1.4 Circumscribing Textual Entailment

Such observations as the ones presented in Section 1.3 regarding the textual entailment datasets employed by the RTE challenges have initiated a wider exchange of opinions in the research community.

Among the first to identify and spotlight the importance of a more mature definition of the textual entailment task, which will allow for clear distinctions among the different types of inferences involved, have been Zaenen et al. (2005). In a critical discussion of the datasets used in RTE-1, Zaenen et al. focus on several of their properties, which they clearly see as weaknesses. These include unnormalized spelling instances, the underspecification of the scope of the required world knowledge, as well as issues regarding human inter-annotator agreement with respect to entailment values.

Characteristically, Zaenen et al. note that the information packaging within the pairs demonstrates a high level of similarity, ignoring the constructional and lexical range that can be used to express an idea. As this will not correspond to the real demands of applications such as QA systems, they propose an augmentation of the types of pairs constructed by existent RTE techniques, with a determined portion that explicitly accounts for the various types of entailments[3].

Manning (2006) supports a different view regarding the issues raised. Sharing similar concerns about the practical usability of the task, but from an opposing perspective, he maintains that an attempt of circumscribing a natural task such as textual entailment recognition might cause degrading

[3] Or local textual inferences, in their terms, which can basically be analyzed in presuppositions, conversational implicatures and "genuine" entailments.

rather than desirable effects. Manning concludes that the use of artificially constructed text may undermine the operational utility of systems and the scientific goals of the challenge.

In response to these claims, Crouch et al. (2006) argue for an approach that will legitimize both naturally occurring text data, and laboratory ones, which will allow for the isolation and identification of core phenomena. They illustrate this by foregrounding the particular attempt in this direction made by the AQUAINT Knowledge-based Evaluation (Crouch et al., 2005), on which we elaborate in Section 2.1.1.

1.5 Overview of the Thesis

The present work makes concrete contributions to this vivid debate about the tenability and practical worth of the textual entailment task by actively investigating tangible ways of creating a better-founded RTE setting. Annotation and evaluation of textual entailment datasets are the main tools in this process.

After a critical review of earlier, preliminary attempts in this direction, in Chapter 2, we introduce in Chapter 3 a new model for evaluation of textual entailment datasets. The model, Annotating RTE (ARTE) scheme, is a scheme for manual annotation of textual entailment—both positive and negative—that makes it possible to pinpoint variant semantic-linguistic properties of entailment in the data.

The annotation of the RTE-2 Challenge dataset based on the ARTE scheme enables a direct analysis of the contribution of individual inference mechanisms in the dataset, and an evaluation of their distribution across its various subsets. The results of this analysis, as well as of a broader examination of the dataset's characteristics, are laid out in Chapter 4.

As Dagan et al. (2006) point out, an analysis of the performance of the existent textual entailment systems, relative to different types of entailments, is likely to bring forth future improvements in textual entailment technology. Chapter 5 attempts exactly such an analysis, by taking into account different factors, including the annotation results of ARTE . It also examines the relationship between systems' performance and the interesting notion of overlap–entailment correlation, which we introduce in Chapter 4.

The main findings of the thesis are summarized in Chapter 6, where ideas for future research are also presented. In the end, Appendix A contains the full version of the ARTE guidelines, exhibiting a large number of textual entailment examples and their annotation.

Chapter 2

Evaluating Textual Entailment Datasets

Chapter 1 introduced the task of textual entailment and foregrounded the need for an extensive and diverse evaluation of the datasets used for its purposes. In this chapter we review such attempts and comment on them.

2.1 Review of Previous Work

The theoretical discussion about the foundations of the textual entailment task presented in Chapter 1 has been complemented with empirical contributions towards a concrete evaluation framework for RTE. The research conducted in this direction, though, has generally been rather fragmentary and of limited scope.

2.1.1 A First Annotation Scenario

As mentioned in Section 1.4, the AQUAINT Knowledge-Based Evaluation (KBEval) Pilot provides an annotation scheme for textual inference (Crouch et al., 2005) that constitutes one of the earliest such attempts. In fact, the scheme is a refinement of the PASCAL gold standard annotation scheme, in that it proposes a three-way classification of pairs according to their entailment value, as well as a number of additional annotation fields. The scheme's main dimensions can be summarized in the following way:

Polarity: *true*, *false* or *unknown*.

It corresponds to the entailment value of the pair. The value *true* indicates positive entailment, while *false* and *unknown* induce a natural partition of the RTE negative entailment set into two categories: the pairs in which T and H contradict each other, and the ones in which H can neither be inferred nor contradicted by T.

Force: *strict* or *plausible*.
It indicates whether additional context could affect the polarity value, aiming at a distinction between certain and plausible inferences.

Source. *linguistic* or *world*.
It characterizes the type of reasoning associated with the entailment, according to whether a competent but ignorant speaker of the language would be in position to judge the polarity.

The scheme additionally offers various optional fields, such as human readable explanations, which have a more experimental status.

2.1.2 Syntactic and Lexical Levels of Entailment

Since the release of the RTE datasets there has been a small number of attempts at defining and analyzing distinct levels or layers of entailment conceived in them. The majority of them focus on phenomena that correspond to well-studied NLP tasks and can be captured by robust tools and resources—namely, syntactic and lexical phenomena.

Vanderwende et al. (2005)

Vanderwende et al. (2005) examine the complete test set of RTE-1 with the purpose of isolating the pairs whose categorization can be accurately predicted based solely on syntactic cues. The syntactic level of entailment defined in this way involves phenomena considered as possible-to-handle exclusively with a typical state-of-the-art parser, and includes argument assignment, intra-sentential pronominal anaphora and several structural alternations. Their human annotation indicates that a portion of 37% of the entailments are decided merely at the syntactic level, while this figure climbs to 49% if the information of a general-purpose thesaurus is additionally exploited.

Bar-Haim et al. (2005)

Bar-Haim et al. (2005) take this idea a step further and annotate 30% of the RTE-1 test set at two strictly defined levels of entailment. Extending Vanderwende et al.'s work, they consider a lexical entailment level, which involves morphological derivations, ontological relations and lexical world knowledge, in addition to a lexical-syntactic level, which, on top of lexical transformations, contains syntactic transformations, paraphrases and coreference.

The annotations are viewed as the classifications made by an idealized system that achieves a perfect implementation of the inference mechanisms regarded. In this system-oriented perspective, the notion of recall is naturally applied for evaluation of success, which yields a recall of 44% for the lexical and of 50% for the lexical-syntactic level. Moreover, Bar-Haim et al. make an evaluation of the distribution of each of the inference mechanisms present at each level, where they report paraphrases and syntactic transformations as the most notable contributors.

Glickman (2006)

Finally, Glickman (2006) defines a lexical reference subtask, which reflects the extent to which the concepts of H are explicitly or implicitly referred to in T. Thus this entailment subtask, or level, can be regarded as a natural extension of textual entailment for sub-sentential hypotheses. Its evaluation in a sample of the RTE-1 development set produces a recall of 69%, suggesting that lexical reference plays an important, but not sufficient, role in RTE.

2.1.3 The Role of Lexical and World Knowledge

With the release of the RTE-3 datasets, Clark et al. (2007) explore the requirements of RTE in a way that differs from the previous approaches in that it is not centered on the basic lexical-syntactic levels of entailment, but instead it investigates a wide range of phenomena involving lexical and world knowledge.

Clark et al. manually annotate 25% of the positive entailment pairs in RTE-3 for thirteen distinct entailment phenomena. Three different, though loosely delineated, types of world knowledge are covered in this compilation: general world knowledge (i.e. nondefinitional facts about the world), core theories knowledge (e.g. space and time) and knowledge related to frames and scripts (i.e. stereotypical situations and events). Some of the other phenomena involve implicative verbs, metonymy and protocol.

The frequency statistics of the sample indicate that the vast majority of entailments require a significant amount of world knowledge, and especially of the general, nondefinitional type. Hence the acquisition of this type of knowledge is one of the most essential requirements that the RTE-3 Challenge poses to participating systems.

2.2 Discussion

The AQUAINT scheme allows for certain forms of system error analysis; e.g., whether a system produces answers that are demonstrably false, or merely unjustified by the data. Still, its distinction of the types of reasoning involved in the data is rather crude, and at the same time notoriously elusive.

Regarding the three approaches of Subsection 2.1.2, though they make valuable explorative contributions to the evaluation of textual entailment, the common denominator among them is that they do not target at a complete and full-fledged analysis of the datasets. Namely, each of them disregards to a different extent certain important aspects of the entailment phenomenon, such as logical inference or deep semantic understanding of the text. As a result, no full coverage of the data is provided.

Apart from that, all three approaches were only attested to the dataset of the first challenge RTE-1, which, as the first germinal construction of the kind, is associated with several idiosyncrasies. The ones noted in Section 1.3 are some of them; the fact that the contribution of coreference in the dataset suffers, since, according to (Dagan et al., 2006), the annotators were instructed to reduce the complexity by replacing anaphors with their appropriate reference, is another.

Finally, the contribution of Clark et al. in the discussion of how an appropriate framework for the evaluation of RTE can be set is significant in that it provides a rich representation of entailment phenomena, some of which extend beyond the traditional syntactic and lexical levels. In particular, the study valuably adds to the investigation of world knowledge, and how it can be specified in RTE.

Nonetheless, like previous studies, (Clark et al., 2007) does not aim at a complete coverage of the entailment phenomena. Its focus on the analysis of world knowledge leaves certain basic inferences of syntactic or grammatical type aside (e.g. coreference resolution, named entity recognition, coordination etc.). Furthermore, the study of negative entailment is left entirely beyond the scope of this work.

In summary, the frameworks discussed in this chapter are important, but rather preliminary attempts at the evaluation of textual entailment datasets.

Chapter 3

The ARTE Scheme

In Chapter 2 we presented previous works on evaluation of textual entailment datasets, and we critically discussed their contributions. In the present chapter we make a new contribution in this area by proposing a unified, comprehensive framework for the evaluation of textual entailment recognition: the Annotating RTE (ARTE) scheme.

In ARTE the datasets are manually annotated and analyzed for a wide variety of entailment phenomena that cover the whole spectrum of local textual inferences. Unlike the ones reviewed in Chapter 2, the annotation scheme presented here does not function selectively on a portion of the various entailment phenomena, but it uniformly accounts for all types of phenomena encountered.

Apart from that, the scheme takes a new perspective on the classification of the entailment phenomena. ARTE views the entailment problem in relation to three well-defined levels: Alignment, Context and Coreference. The potential of each level is explored in depth for the positive entailment cases, while in the negative ones we aim at a more basic, elementary scheme that allows for solid observations on the particularities of non-entailment.

In the following sections we briefly introduce the basic definitions and main ideas associated with each component of the architecture. A more profound view, with additional details and strict guidelines regarding their use, is available in Appendix A. The technical aspects of the project are also elaborated there.

3.1 The Scheme for Positive Entailment

The main concept behind the annotation scheme for the positive entailment pairs is that of alignment, in a sense similar to what RTE systems typi-

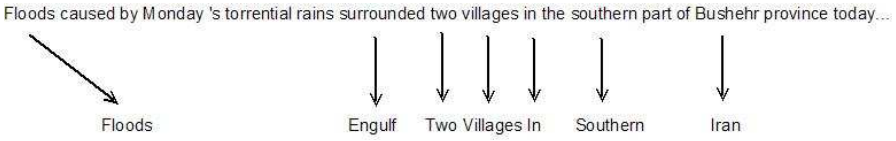

Floods caused by Monday 's torrential rains surrounded two villages in the southern part of Bushehr province today...
Floods
Engulf
Two Villages In
Southern
Iran

Figure 3.1: Alignment produced by the system of Bayer et al. (2005), which treats entailment data as an aligned translation corpus, and bases its judgment on translation quality measures. The system induces alignment models using the GIZA++ toolkit (Och and Ney, 2003).

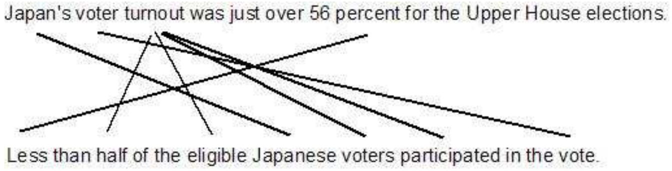

Japan's voter turnout was just over 56 percent for the Upper House elections.
Less than half of the eligible Japanese voters participated in the vote.

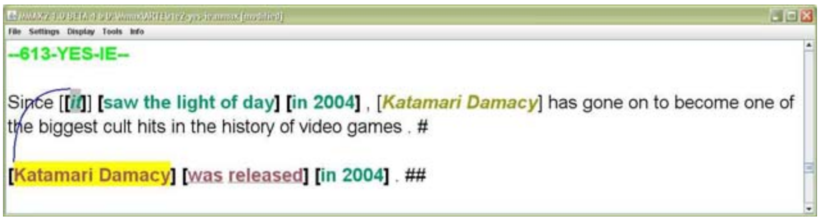
Figure 3.2: The probabilistic setting of Glickman et al. (2005) induces an alignment between the terms of T and H in a way similar to alignment models in statistical MT.

cally use in order to model entailment. The idea is roughly illustrated in Figures 3.1 and 3.2, which represent outputs of such systems.

Our view of alignment is grounded on this underlying idea; however it explicitly considers alignment at a level beyond bags of words. It largely takes syntactic structure into account, and models the task carried out by syntactic matching systems, as the bag-of-words scheme does for lexical systems. It could thus be regarded as a “flat” approximation of a graph subsumption model.

The direction of the alignment is, similarly to (Wang and Neumann, 2007), from H to T, so that H is covered exhaustively while T may contain irrelevant parts that are not aligned. However, unlike the automatic system outputs, the type of alignment presented here is a manual human construction.

Figure 3.3 exemplifies this with a T–H pair from the RTE-2 test set: The alignment of the subject *Katamary Damacy* in H is produced in a way that respects the matching of the complete syntactic structure of H, as it points to the corresponding subject, and not to the lexically identical phrase of T. The anaphoric coreference relation present in T is captured individually at the Coreference level. In parallel a third level, Context—not active in this example—, models the contribution of higher-level factors



--613-YES-IE--

Since [[*it*]] [saw the light of day] [in 2004] , [*Katamari Damacy*] has gone on to become one of the biggest cult hits in the history of video games . #

[Katamari Damacy] [was released] [in 2004] . ##

Figure 3.3: In ARTE alignment is guided by syntactic structure, approximating a graph subsumption model.

typically outside the boundaries of the local syntactic structures, such as factivity and polarity.

The relevant fragments of T and H selected and annotated at any of these levels are called *markables*, and tend to roughly correspond to basic syntactic constituents, although this is not a formal requirement and there are numerous divergencies. Each level of the annotation is related to its own set of markables, which may be discontinuous and/or overlapping.

3.1.1 Alignment

The Alignment level is intended to capture basic inherent properties of the textual entailment phenomenon and takes up a twofold function. On the one hand, it provides directed pointer relations (alignments) from the constituents of H to the corresponding parts of T that are responsible for the local entailment. On the other hand, it provides information about the specific nature of the alignment constructed between the two. Every such alignment is associated with a label that describes it and indicates what type of textual inference has made it possible.

In total there are ten different features serving as labels for this purpose, two of which are further refined in subcategories. Moreover, the features are not mutually exclusive, but can be applied in combinations so as to achieve a result as informative as possible. The list of features is as follows:

Identity indicates that the alignment roughly involves a mere surface-level lexical match of the two markables. What is meant by this is not strict string equality, but rather similarity that allows for minor variations (e.g. tense[1], inflection or different prepositions) that do not have sig-

[1] In fact we follow one of the guidelines presented in (Dagan et al., 2006), and ignore tense aspects entirely, as T and H may originate from documents at different points in time.

nificantly different semantic interpretations in the particular context. Figure 3.4 provides one such example.

Furthermore, surface similarity is a necessary but not sufficient condition for an Identity annotation: If another, more specific feature is applicable, it will be the one selected.

Coreference indicates that the markables aligned are coreferent. Typically the H-markable involved is a noun phrase (NP), while the T-markable either an NP or a pronoun/relativizer. Hence this feature is not restricted to the anaphoric type of coreference. Figure 3.3 provided an example.

Genitive marks an alignment that is based on the analysis of genitive case, signaled by a possessive pronoun, the possessive clitic *'s*, or the preposition *of*. It involves, therefore, the matching of a semantically underspecified construction, which can denote a number of different relations (e.g. alienable/inalienable possession, composition, origin, etc.), to the specific interpretation it acquires in a particular context. Figure 3.5 presents an example.

Modifier indicates that the alignment relies on the direct interpretation of a modifier—either adjectival (or nominal, in the case of compound noun constructions) or adverbial—, which, similarly to the Genitive case, explicates an unspecific relation. An example is provided by Figure 3.6.

Morphological applies in case the alignment represents a morphological[2] transformation. Considered in this category are only word-formation rules, and not inflectional rules—inflectional variations do not weigh heavily in our textual entailment framework and are typically modeled with Identity alignments.

There are four distinct subcategories to specify the particular type of transformation.

Nominalization. The aligned pair consists of a verb- or adjective-markable, and a derivationally related noun-markable; e.g., *make* ⟷ *maker*.

[2] Our rather broad use of the term *morphological* here is conventional, mainly serving to group together several closely related mechanisms. For example, it encompasses cases of nominalization which do not in fact affect morphology, such as *purchases* (N) ⟷ *purchases* (V). Therefore the sense defined here should not be related to linguistic debates about the nature of morphology and its rules.

Demonym. The alignment involves a place and its inhabitants, or a people and its members; e.g., *Liberia* ⟷ *Liberian.*

Acronym. The alignment involves a phrase and its typical abbreviation, formed by the initial letters or parts of its words; e.g., *New Jersey* ⟷ *N.J.*.

Other. Any other type of non-inflectional morphological transformation; e.g., *big* ⟷ *the biggest*, *random* ⟷ *randomly.*

Argument Variation marks an alignment between two predicates with variation in their argument structure, i.e., realizing corresponding arguments using different grammatical functions, as in Figure 3.7.

Passivization. This label can be applied to Argument Variation alignments that are between predicates appearing in different grammatical voices (active and passive); e.g., *killed* ⟷ *was killed.*

Ontological indicates that the alignment involves one of the most common lexical ontological relations, mainly drawn from the lexical semantic resource of WordNet (Fellbaum, 1998). The relations chosen are the ones that are typically associated with the notion of *semantic similarity*: synonymy and hypernymy.

Synonymy indicates that the two markables are interchangeable within the context in which they appear; e.g., *a human* ⟷ *a human being.*

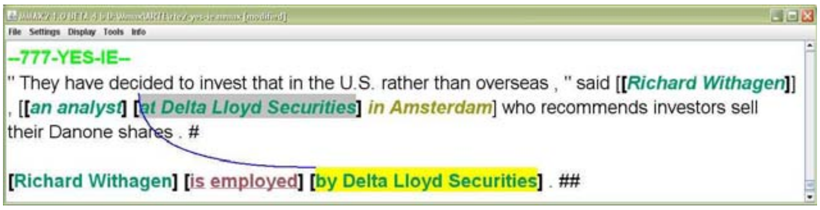
Hypernymy indicates that the two markables are linked by the is-a-kind-of relation; e.g., *spokeswoman* ⟷ *representative.*

Quantities marks an alignment which involves reasoning based on quantities and quantifiers, as in Figure 3.8.

Reasoning. This final feature is rather the most comprehensive, as it encompasses all cases that extend beyond the rest of the features, and represents several different forms of reasoning.

These may involve a *lexical relation* not among the aforementioned ones, *general world knowledge*, *geographical*, *spatial* or *temporal knowledge*, *modality* markers, *punctuation*, *logical* or other *general inference mechanisms*, *metonymy*, *elliptical constructions*, *conversational implicatures* or indirect contributions of the sentences' *context.*

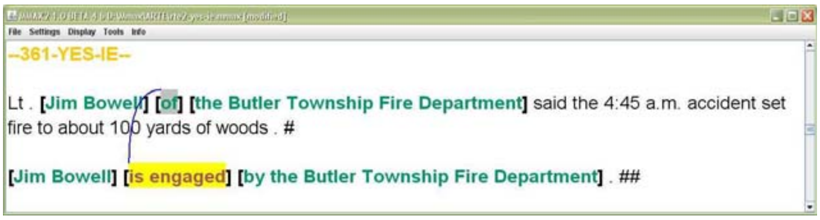
Figures 3.9 and 3.10 present such cases; Appendix A contains a larger number of examples.



--777-YES-IE--
" They have decided to invest that in the U.S. rather than overseas , " said [[Richard Withagen]] , [[an analyst] [at Delta Lloyd Securities] in Amsterdam] who recommends investors sell their Danone shares . #

[Richard Withagen] [is employed] [by Delta Lloyd Securities] . ##

Figure 3.4: The two markables are in an Identity alignment relation, although they are not lexically identical.

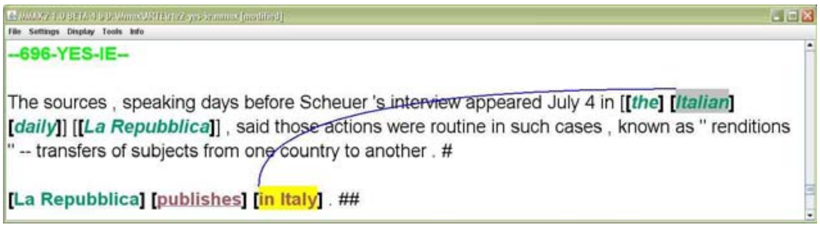


--361-YES-IE--

Lt . [Jim Bowell] [of] [the Butler Township Fire Department] said the 4:45 a.m. accident set fire to about 100 yards of woods . #

[Jim Bowell] [is engaged] [by the Butler Township Fire Department] . ##

Figure 3.5: An alignment labeled as Genitive, as the phrase *is engaged* in H specifies the relation indicated with the preposition *of* in T.

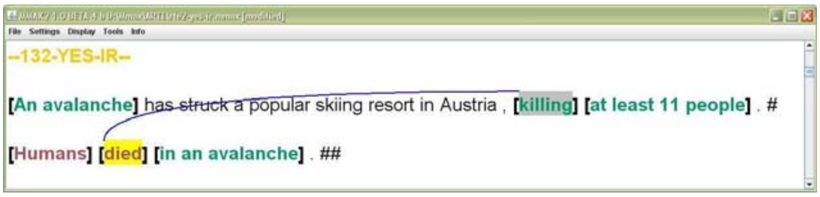


--696-YES-IE--

The sources , speaking days before Scheuer 's interview appeared July 4 in [[the] [Italian] [daily]] [[La Repubblica]] , said those actions were routine in such cases , known as " renditions " -- transfers of subjects from one country to another . #

[La Repubblica] [publishes] [in Italy] . ##

Figure 3.6: An alignment with a Modifier label. The modifier in H *Italian* in this case denotes location. Note that this is also an instance of Demonym.

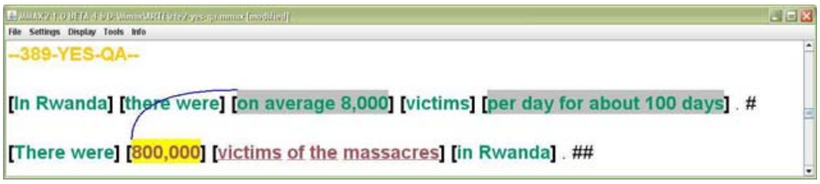


--132-YES-IR--

[An avalanche] has struck a popular skiing resort in Austria , [killing] [at least 11 people] . #

[Humans] [died] [in an avalanche] . ##

Figure 3.7: An alignment marked as Argument Variation, since the arguments of the two predicates are aligned by different syntactic functions.

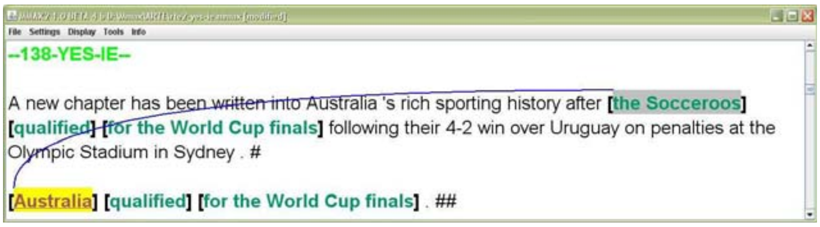


--389-YES-QA--

[In Rwanda] [there were] [on average 8,000] [victims] [per day for about 100 days] . #

[There were] [800,000] [victims of the massacres] [in Rwanda] . ##

Figure 3.8: An alignment marked as Quantities, since it requires arithmetic reasoning.

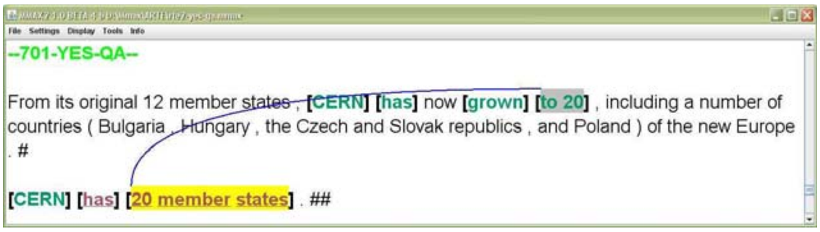


--138-YES-IE--

A new chapter has been written into Australia 's rich sporting history after [the Socceroos] [qualified] [for the World Cup finals] following their 4-2 win over Uruguay on penalties at the Olympic Stadium in Sydney . #

[Australia] [qualified] [for the World Cup finals] . ##

Figure 3.9: This alignment, labeled with Reasoning, is grounded on reasoning involving the figure of speech of antonomasia.



--701-YES-QA--

From its original 12 member states , [CERN] [has] now [grown] [to 20] , including a number of countries (Bulgaria , Hungary , the Czech and Slovak republics , and Poland) of the new Europe . #

[CERN] [has] [20 member states] . ##

Figure 3.10: This alignment is due to the analysis of the NP-ellipsis in T *20*, and is marked as Reasoning.

3.1.2 Context

In certain cases, even if there is a prefect alignment between T and H, the entailment may still not hold, due to the interference of external context factors that block it. For instance, as Nairn et al. (2006) point out, there are clear semantic differences among sentences such as

(1) a. Ed closed the door.
b. Ed did not close the door.

(2) a. Ed forgot that he had closed the door.
b. Ed did not forget that he had closed the door.
c. Ed forgot to close the door.
d. Ed did not forget to close the door.

(3) a. Ed claimed that he had closed the door.
b. Ed did not claim that he had closed the door.
c. Ed pretended that he had closed the door.
d. Ed did not pretend that he had closed the door.

Obviously the truth of statement (1a) follows from the truth of both the assertion (2a) and its negation (2b). On the other hand, neither of these inferences can be drawn from the truth of (3a) or its negation (3b). Moreover, both (3c) and (3d) entail statement (1b). Finally statement (2c) is of a special nature, since it entails the negation (1b), while its appearance in a negative polarity context as in (2d) entails the assertion (1a).

The Context level of the scheme is designed to provide such information about the relative polarity forced by the context in which statements are made, and which may change their interpretation, or the author's commitment to them. This problem is deeply related to the one of assessing the veridicity of textual content, which is an issue of high importance for textual inferences. An interesting discussion on this delicate topic is given by Karttunen and Zaenen (2005).

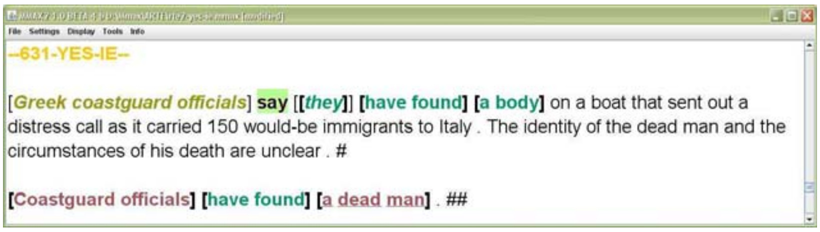
An example of Context annotation is presented in Figure 3.11. The annotation at this level is focused on the contents of T outside the boundaries of the aligned markables. In particular it involves the following two features:

Factivity measures the degree of the author's commitment to the truth of the statement in the complement clause introduced by the markable. It is assigned one of four possible values.

- **Neutral** is selected in case the complementizer carries neither presuppositions nor entailments and therefore does not impose any commitment to the truth/falsity of the subordinate clause. This would apply for example in contexts of report, belief, volition, planning or commission; e.g., *say*, *claim*.
- **Factive** indicates that the truth of the complement clause is presupposed; e.g., *reveal*, *uncover*.
- **Counterfactive.** In contrast to the Factive case, this value indicates that the falsity of the complement is presupposed. A typical predicate in this category is *pretend*.
- **Implicative.** This category comprehends the three subgroups of two-way implicatives, one-way +implicatives and one-way –implicatives, as presented in (Nairn et al., 2006). It indicates that the complementizer carries entailments and possibly also presuppositions, and the former may change if the relative polarity of the sentence changes.

  Typical examples of implicative expressions are predicates such as *manage*, *refuse* and *attempt*, when they introduce nonfinite complements.

Negation signals that the markable imposes a negative polarity on the complement clause. Carriers of such information may be the negation particle *not*, the downward-monotone quantifier *no*, restricting prepositions such as *without* and *except*, or certain subordinating conjunctions such as *unless*.



--631-YES-IE--

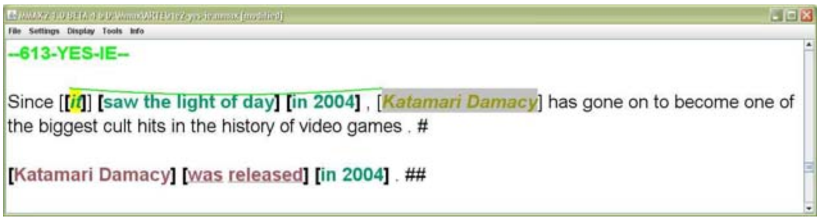
[*Greek coastguard officials*] **say** [[*they*]] [**have found**] [**a body**] on a boat that sent out a distress call as it carried 150 would-be immigrants to Italy . The identity of the dead man and the circumstances of his death are unclear . #

[**Coastguard officials**] [**have found**] [**a dead man**] . ##

Figure 3.11: In this pair the information about the event described in H lies in a report context, introduced by the neutral factivity predicate *say*. The entailment holds only if we trust in the veridicity of the source of this information.

3.1.3 Coreference

As remarked in the introduction of Section 3.1, the Coreference Level is designed to provide an additional layer of information in those cases of coreference, in which resolution is crucial for the entailment. Figure 3.12 demonstrates how this is achieved for the pair of Figure 3.3.



--613-YES-IE--

Since [[*it*]] [saw the light of day] [in 2004] , [*Katamari Damacy*] has gone on to become one of the biggest cult hits in the history of video games . #

[Katamari Damacy] [was released] [in 2004] . ##

Figure 3.12: At the Coreference level the pronoun of T *it* is linked to its antecedent *Katamari Damacy*.

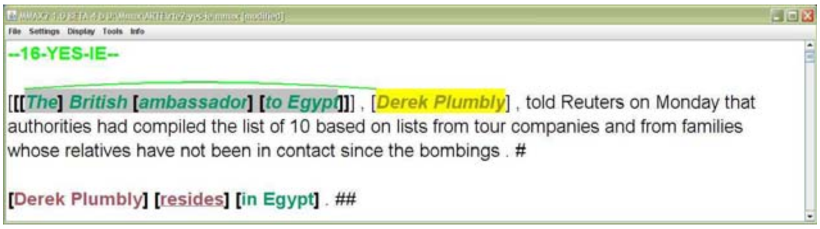
A number of different types of coreference is captured in this way, providing a rich account of coreference mechanisms from several perspectives. Similarly to alignment relations, coreference relations link together two different markables and each such link is labeled with one of the following features:

Supplemental applies to a coreference that involves an NP and an expression supplemental to it. Two distinct subcategories refine this idea:

- **Apposition** indicates that the two coreferent parts are in an apposition construction, as in Figure 3.13.
- **Reduced Relative** captures the relation between an NP and a reduced relative clause modifying it, as in Figure 3.14. Obviously, since the relative pronoun is missing, this feature expresses a coreference not directly evident, but rather implicit in the grammatical analysis of the construction.

Anaphoric marks an explicit coreference of anaphoric type; this is further specified with the two following subcategories:

- **Pronominal** indicates that there is a coreference between the markables, established by the reference resolution of a pronoun (e.g., relative, personal, demonstrative, possessive). Figure 3.12 above provided an example of pronominal coreference.
- **Nominal** applies in case two NP-markables are coreferent without being in a direct syntactic relation (e.g. appositive or equative constructions), as in Figure 3.15.

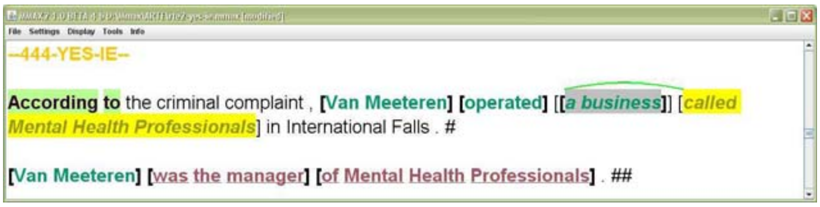


--16-YES-IE--

[[[The] British [ambassador] [to Egypt]]] , [Derek Plumbly] , told Reuters on Monday that authorities had compiled the list of 10 based on lists from tour companies and from families whose relatives have not been in contact since the bombings . #

[Derek Plumbly] [resides] [in Egypt] . ##

Figure 3.13: The NP *Derek Plumbly* in T is a supplemental expression of the appositive type to the NP *The British ambassador to Egypt*.

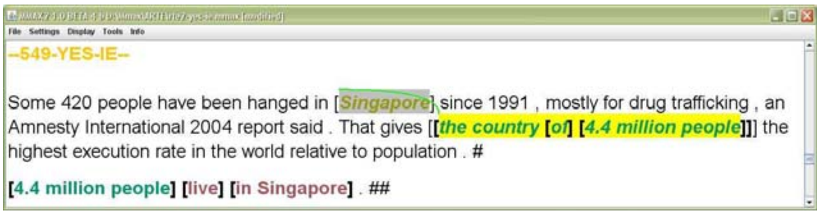


--444-YES-IE--

According to the criminal complaint , [Van Meeteren] [operated] [[a business]] [called Mental Health Professionals] in International Falls . #

[Van Meeteren] [was the manager] [of Mental Health Professionals] . ##

Figure 3.14: The NP *a business* is linked to the reduced relative clause *called Mental Health Professionals*, to which it functions as subject.



--549-YES-IE--

Some 420 people have been hanged in [Singapore] since 1991 , mostly for drug trafficking , an Amnesty International 2004 report said . That gives [[the country [of] [4.4 million people]]] the highest execution rate in the world relative to population . #

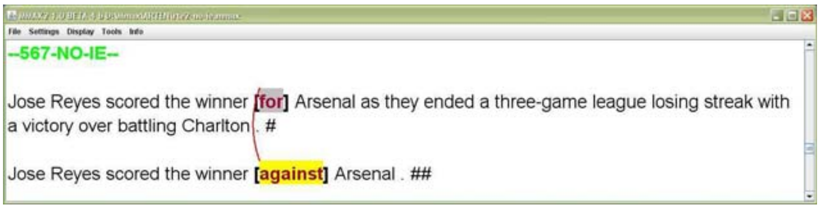
[4.4 million people] [live] [in Singapore] . ##

Figure 3.15: An example of Anaphoric Coreference of type Nominal between the two NPs.

3.2 The Scheme for Negative Entailment

Negative entailment detection imposes different kinds of challenges on systems, as the task of pinning down the reasons for absence of an entailment relation can be far more evasive and subtle than the one of highlighting the existent evidence for its presence. For this reason our negative entailment scheme has a status more experimental and less analytical than the fully-developed positive one.

In this explorative setting we aim at a classification of the negative entailment cases into three major categories, according to the most prominent and direct reason why the entailment cannot be established. Though in many cases there are several small pieces of evidence for non-entailment, we focus on the single one that we consider as the most obvious "trap" for systems (and humans) judging the entailment. Figure 3.16 presents a typical non-entailment annotation.



--567-NO-IE--

Jose Reyes scored the winner [for] Arsenal as they ended a three-game league losing streak with a victory over battling Charlton . #

Jose Reyes scored the winner [against] Arsenal . ##

Figure 3.16: The prepositions *for* and *against* in this pair convey diametrically different meanings that contradict each other.

The categories defined involve context, additional information and misalignment factors, and are as follows:

Context indicates that the entailment is blocked by the presence of a particular context which modifies the truth value of the rest. This context may involve for instance modality, a restricted spatiotemporal frame, negation, non-factivity, or an expression affecting the relative polarity of its complement. An example is given in Figure 3.17.

Additional indicates that H is more informative than T. On the one hand H might possibly be partially entailed by T, but on the other hand there is additional information present in H which cannot be inferred from T. Figure 3.18 presents an example.

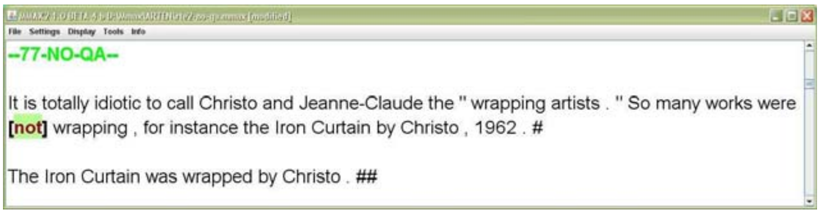
Misalignment suggests that H is partially aligned to T in a way that respects the entailment, but the remaining part of H aligns to a part of

T that is either inadequate for the entailment, or even contradictory. It is refined by means of the following two subcategories:

Inadequacy. The misalignment is specified as Inadequacy in case the information available in T is insufficient to support the corresponding information conveyed by the misaligned H-markable. This means that H as such could be true given T, but its truth is not assured by the truth if T, as Figure 3.19 illustrates.

Contradiction. Finally, the misalignment is labeled as Contradiction in case it proves not merely that H is not entailed by T, but also that the two in all likelihood cannot be both true at the same time, if interpreted in exactly the same spatiotemporal frame, referring to the same events/situations.

This definition follows the annotation guidelines for marking contradictions (Manning et al., 2007) introduced in the RTE-3 Optional Pilot Task: Extending the Evaluation of Inferences from Texts[3]. Figure 3.16 is one such example.

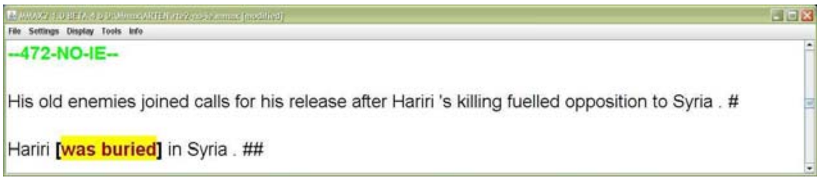


--77-NO-QA--

It is totally idiotic to call Christo and Jeanne-Claude the " wrapping artists . " So many works were [not] wrapping , for instance the Iron Curtain by Christo , 1962 . #

The Iron Curtain was wrapped by Christo . ##

Figure 3.17: The negation particle *not* is blocking the entailment in this pair, which would otherwise hold. Therefore it is annotated with the Context feature.



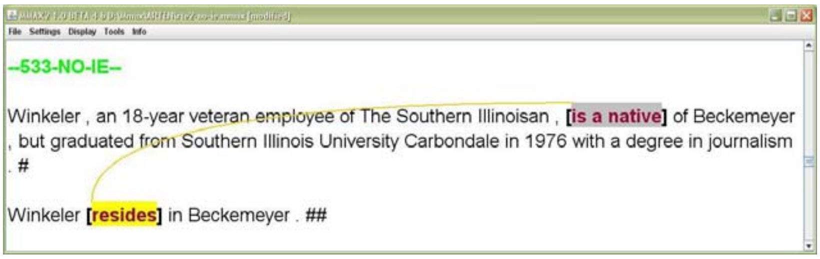
--472-NO-IE--

His old enemies joined calls for his release after Hariri 's killing fuelled opposition to Syria . #

Hariri [was buried] in Syria . ##

Figure 3.18: In this pair T contains no information related to the predicate of H *was buried*. Therefore the latter is marked as Additional.

[3] http://nlp.stanford.edu/RTE3-pilot



--533-NO-IE--

Winkeler , an 18-year veteran employee of The Southern Illinoisan , [is a native] of Beckemeyer , but graduated from Southern Illinois University Carbondale in 1976 with a degree in journalism . #

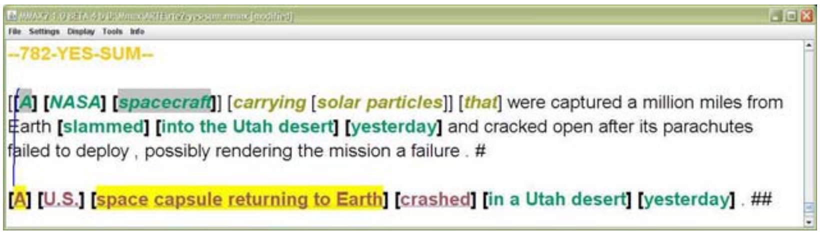
Winkeler [resides] in Beckemeyer . ##

Figure 3.19: The predicate of H is misaligned to a T-markable whose semantic interpretation cannot justify an entailment. Hence the Misalignment is labeled as Inadequacy.

3.3 Discussion

The annotation scheme for RTE presented seeks a balance between richness on the one hand, and usability, on the other. The main principle guiding the selection and definition of features was to create a scheme which is expressive enough to enable useful insights into the data, but at the same time functional enough to allow for a clear and consistent effect. The philosophy served was the one of achieving a "good" trade-off between linguistic sophistication and practicability.

Weighing the strengths and weaknesses of ARTE in this light, we cannot fail to consider the inherent limitations of the "flat" alignment model we adopted. As a concrete example, it is obvious that such a type of alignment is insufficient for sentences with embedded clauses containing information relevant for the entailment, as in Figure 3.20.



--782-YES-SUM--

[[A] [NASA] [spacecraft]] [carrying [solar particles]] [that] were captured a million miles from Earth [slammed] [into the Utah desert] [yesterday] and cracked open after its parachutes failed to deploy , possibly rendering the mission a failure . #

[A] [U.S.] [space capsule returning to Earth] [crashed] [in a Utah desert] [yesterday] . ##

Figure 3.20: An example of a rather "artificial" alignment that is forced by the syntactic structures of the sentences but does not directly reflect the semantic interpretations of the markables involved. A large portion of the information used for this alignment comes from the embedded relative clause in T, which does not participate in the alignment.

As a result, the entailment rules constructed by means of the alignment relations may in certain cases not be self-contained, but count largely on external context factors. This, however, is a natural restriction, since any local entailment relies on its wider context to some extent, and any RTE system must take this into account. Furthermore, ARTE provides explicit indications of such types of contextual alignments by means of the alignment labels (e.g., Coreference, Reasoning) or the—likewise labeled—separate Coreference links.

A conclusion thus drawn is that the ARTE scheme is indeed applicable to textual entailment data, and suitable for modeling them at a generally satisfactory level of success. Regarding the negative entailment scheme, in particular, the resulting annotation suggests that non-entailment pairs can effectively be classified into three major categories. Contrary to what might be expected judging by the intricacy of the phenomenon, these categories have in practice been surprisingly easy to distinguish.

A noteworthy observation is that the annotation of non-entailment provides an indication about the distribution of the non-entailment pairs according to their polarity value, in the sense of Crouch et al. (2005). The binary classification of the negative entailment dataset with respect to polarity is directly reflected in the annotation: Pairs annotated as Contradiction correspond to polarity of value *false*, while pairs annotated as Inadequacy or Additional correspond to polarity of value *unknown*. Context annotation is rather ambiguous with respect to polarity, as pairs marked with it may correspond to either of the two values. It would hence be interesting to investigate a partition of the Context category.

Finally, drawing an analogy between ARTE and the framework of Clark et al. (2007), it is remarkable that the two approaches parallel each other in many points, even though they have been developed entirely independently. Particularly the annotation features comprising each framework are highly comparable, as is the philosophy of investigating entailment beyond the lexical-syntactic level.

On the other hand, of course, this work is in several aspects different from the one of Clark et al.: ARTE presents a model of explicit alignment guided by strict principles, and it handles both positive and negative entailment cases. It was also attested to a larger corpus, consisting of 500 entailment pairs. As a further difference, Clark et al. present a highly fine-grained classification of entailment types that involve world knowledge and reasoning; ARTE is more oriented towards achieving a wide coverage of different types and lacks such a degree of detail in this particular category.

Chapter 4

Analysis of the RTE-2 Dataset

The annotation scheme presented in Chapter 3 enables a direct analysis of the contribution of individual inference mechanisms in textual entailment datasets, and a precise evaluation of their distribution across the diverse subsets associated with different application settings.

In this and the following chapter we lay out the results of such an analysis, conducted on the textual entailment dataset provided by the RTE-2 Challenge. Chapter 4 examines the dataset in its own right, while Chapter 5 carries out a performance-sensitive analysis, exploring the data with respect to the performance of different textual entailment systems on them.

In the present chapter we first prepare the ground for this linguistically oriented analysis by examining the data from a shallow perspective, regarding surface measures such as length of sentences and word overlap. After gaining first insights into the datasets with these observations, we proceed to a deeper evaluation considering the linguistic phenomena behind the given entailments.

4.1 Shallow Measures

A brief inspection of the submission results and systems description of RTE-2, as in (Bar-Haim et al., 2006), reveals that even naive overlap-based systems are able to achieve results comparable to—and sometimes better than—more linguistically sophisticated systems. This observation motivates a shallow-perspective investigation of the datasets.

We examine the complete RTE-2 dataset with regard to the statistical lexical features of sentence lengths and word overlap. The dataset, consisting of 1600 T–H pairs, is split into several subcorpora: The development and test set comprise 800 pairs each, further divided into two halves (400 pairs each) that correspond to the positive and negative entailment instances. The pairs are equally distributed among the four application settings of IE, IR, QA and SUM.

4.1.1 Lengths

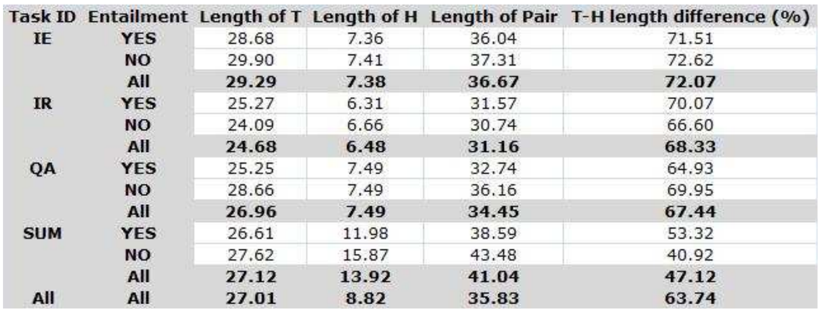
Sentence lengths of the pairs are computed separately for T and H, and the sum of the two defines the length of the pair as a whole. The notion of sentence length stands here for the total number of words that comprise a text snippet; each individual word, including function words, is counted.

The difference of lengths between T and H of each pair is also considered, normalized by the length of T as follows:

$$\text{T–H length difference} = \frac{\text{length of T} - \text{length of H}}{\text{length of T}} \tag{4.1}$$

Table 4.1 presents the statistical distribution of sentence lengths and length differences between T and H across the individual RTE-2 subsets, while Figures 4.1 and 4.2 display the overall average lengths and length differences of pairs for each of the four application settings.

Clearly the tasks demonstrate significant variance in these values, which is the largest between IR and SUM for lengths, and between IE and SUM for length differences. SUM has the longest pairs with approx 41 words on average, and with the smallest difference between the lengths of T and H. On the other hand IR has the shortest pairs, with approx 31 words on



Task ID	Entailment	Length of T	Length of H	Length of Pair	T-H length difference (%)
IE	YES	28.68	7.36	36.04	71.51
	NO	29.90	7.41	37.31	72.62
	All	**29.29**	**7.38**	**36.67**	**72.07**
IR	YES	25.27	6.31	31.57	70.07
	NO	24.09	6.66	30.74	66.60
	All	**24.68**	**6.48**	**31.16**	**68.33**
QA	YES	25.25	7.49	32.74	64.93
	NO	28.66	7.49	36.16	69.95
	All	**26.96**	**7.49**	**34.45**	**67.44**
SUM	YES	26.61	11.98	38.59	53.32
	NO	27.62	15.87	43.48	40.92
	All	**27.12**	**13.92**	**41.04**	**47.12**
All	**All**	**27.01**	**8.82**	**35.83**	**63.74**

Table 4.1: The distribution of average sentence lengths and length differences, as defined in (4.1), across the RTE-2 subsets. Both development and test set are covered.

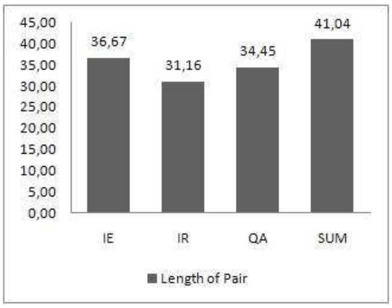


Figure 4.1: The chart of average sentence lengths of pairs for the four tasks.

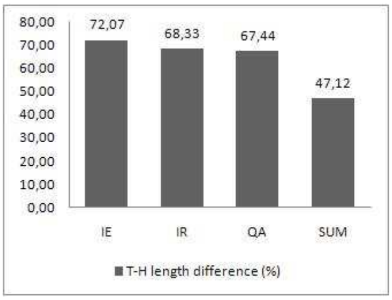
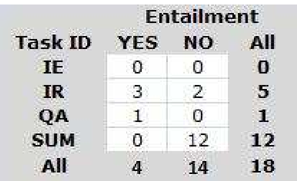


Figure 4.2: The chart of average length differences between T and H of pairs for the four tasks.

average, while IE has the pairs with the greatest length differences, where T is on average by 72.07% longer then H.

Furthermore, differences are also observed between positive and negative entailment pairs. In SUM, for instance, T is longer than H by 53.32 % on the positive entailment pairs, but only by 40.92% on the negative ones.

Finally, Table 4.2 presents the distribution of pairs with negative T–H length difference; i.e., the pairs whose H has greater length than T. The figures are particularly interesting: Although in the RTE task T is typically longer than H—since textual entailment implicates that T contains at least as much information as H, and usually more—, the length of H exceeds the one of T in approx 1.13% (18 out of 1600 pairs) of the data. This "peculiarity" is most common in SUM, and especially in the negative entailment subset.



	Entailment		
Task ID	YES	NO	All
IE	0	0	0
IR	3	2	5
QA	1	0	1
SUM	0	12	12
All	4	14	18

Table 4.2: The distribution of pairs with H longer than T, i.e., negative T–H length difference, as defined in (4.1).

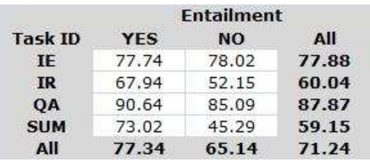
4.1.2 Overlap

The word overlap measure takes into account exclusively content words—namely nouns, non-auxiliary verbs, adjectives and adverbs—and only the base form of those, according to the part-of-speech tagging and lemmatization of the TreeTagger tool (Schmid, 1994). It indicates the relative number of words in H that also appear in T:

$$\text{Overlap} = \frac{\#\ \text{common lemmas of T-H}}{\#\ \text{lemmas of H}} \tag{4.2}$$

The statistical distribution of word overlap across the different application settings is presented in Table 4.3, while Figure 4.3 shows a chart for the positive and negative entailment cases.

We observe that overlap is as high as 71.24% on average for the complete dataset, whereas it varies to different degrees across tasks, and across entailment values within tasks. In particular, entailment value appears to have a highly significant effect on overlap for IR and SUM, but not for IE or QA.



	Entailment		
Task ID	**YES**	**NO**	**All**
IE	77.74	78.02	**77.88**
IR	67.94	52.15	**60.04**
QA	90.64	85.09	**87.87**
SUM	73.02	45.29	**59.15**
All	**77.34**	**65.14**	**71.24**

Table 4.3: The distribution of average word overlap across the four application settings, according to the definition (4.2).

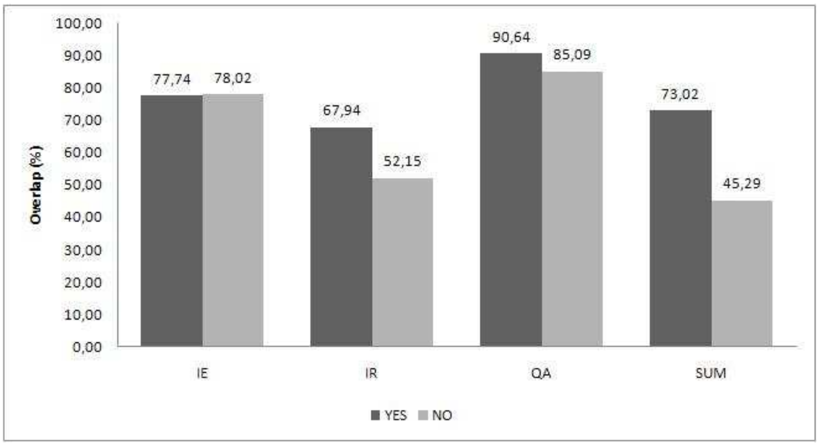


Figure 4.3: The chart of average word overlap in the eight subcorpora of RTE-2.

4.1.3 Overlap–Entailment Correlation

The figures presented in Subsection 4.1.2 motivate a more thorough examination of the correlation between word overlap and entailment value in the datasets. A question arises naturally: To what extent can word overlap serve as a direct indicator of the pairs' entailment value? How well do the datasets fit to a pattern that associates low overlap with non-entailment and high overlap with positive entailment? In other words, how much success do the datasets allow for naive systems that base their judgment only on overlap criteria?

To address this question we define an overlap–entailment correlation measure as follows:

$$\text{Correlation} = \begin{cases} \frac{\text{Overlap}}{100} & \text{if Entailment} = \text{YES} \\ \\ \frac{100-\text{Overlap}}{100} & \text{if Entailment} = \text{NO} \end{cases} \tag{4.3}$$

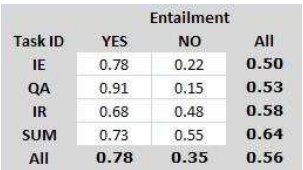
Thus correlation ranges over the closed interval 0 to 1 and is analogous to overlap in case entailment holds; in case of negative entailment it takes lower values when overlap is high, and higher values for lower overlap. In this sense it provides a measure of how successfully the entailment value of a pair can be determined merely on the basis of word overlap.

Table 4.4 presents the distribution of average correlation across the RTE-2 datasets, and Figures 4.4 and 4.5 display this distribution for the different subcorpora. We observe that the values range from 0.50 to 0.64, with an overall average of 0.56. This fact indicates a certain general "favorableness" towards overlap, with the only exception of IE, which demonstrates impartiality. SUM appears to be the most "promising" task for overlap-based methods.

We come back to these results in Subsection 5.1.2 of Chapter 5, where they are interestingly parallelized to systems' performance and further discussed.

4.2 Deep Measures

The evaluation of the datasets with respect to the shallow measures of Section 4.1 sets the stage for a deeper-perspective evaluation, based on semantic-linguistic aspects of entailment. The ARTE annotation scheme presented in Chapter 3 is put to good use for this purpose.



	Entailment		
Task ID	YES	NO	All
IE	0.78	0.22	**0.50**
QA	0.91	0.15	**0.53**
IR	0.68	0.48	**0.58**
SUM	0.73	0.55	**0.64**
All	**0.78**	**0.35**	**0.56**

Table 4.4: The distribution of average overlap–entailment correlation across the four application settings, according to the definition (4.3).

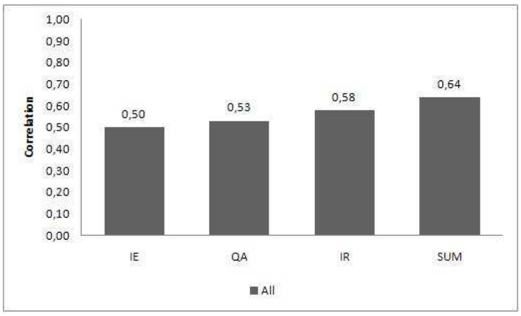


Figure 4.4: The chart of average overlap–entailment correlation in the four application settings.

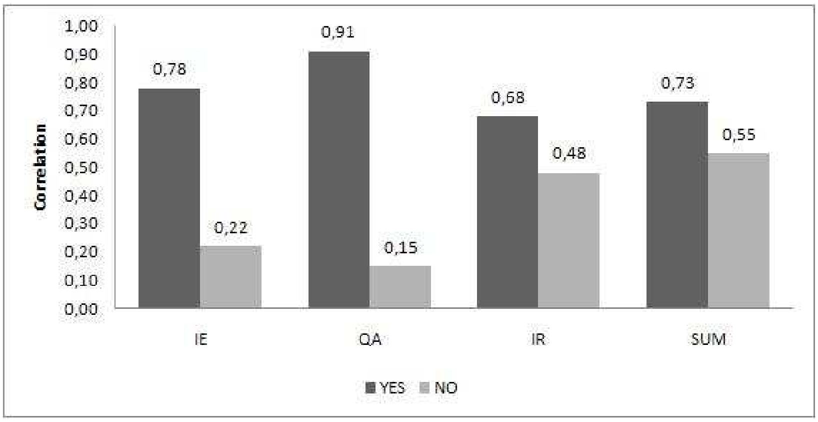


Figure 4.5: The chart of average overlap–entailment correlation in the different RTE-2 subcorpora.

The ARTE scheme was applied to the complete positive entailment test set of RTE-2 (400 pairs; i.e. 100 pairs of each task), as well as to a random 25% portion of the negative entailment test set, equally distributed among the four tasks (100 pairs; i.e. 25 pairs of each task). An overall 62.50% of the RTE-2 test set was thus covered.

Two human annotators worked independently on the annotation, frequently meeting for adjudication of disagreements. Any remaining points of conflict were cleared by a third annotator. As the annotation guidelines were under development throughout the process, inter-annotator agreement figures were not computed.

Admittedly, this was a decision that deprived the annotation from the benefits of a direct quality assessment measure. We are nonetheless confident that this weakness is outweighed by the advantage of the carefully developed and comprehensive set of annotation guidelines produced in this way.

The guidelines are presented in Appendix A, while the annotation itself is planned to be made publicly available in the near future. In this section we lay out the statistical analysis results of the annotation separately for the positive and negative entailment case.

4.2.1 The Positive Entailment Dataset

ARTE provides us with a total of 23 different features for positive entailment annotation. One of them, Identity, is of a special nature, since it stands rather for lack of other features than as a feature in its own right. Additionally, the feature Counterfactive counts zero occurrences in the annotated dataset. Therefore the features Identity and Counterfactive are sometimes disregarded in the following analysis.

The annotated features offer several ways to make useful classifications of the data.

Individual Entailment Features Distribution

The most straightforward observations come from the simple frequency counts of the occurrences of each individual entailment feature in the datasets. For this purpose we ignore multiple occurrences of a feature in a certain pair. Thus the frequencies presented indicate the number of T–H pairs that have been labeled with the feature in question for at least one of the constructed alignments from H to T of the given pair.

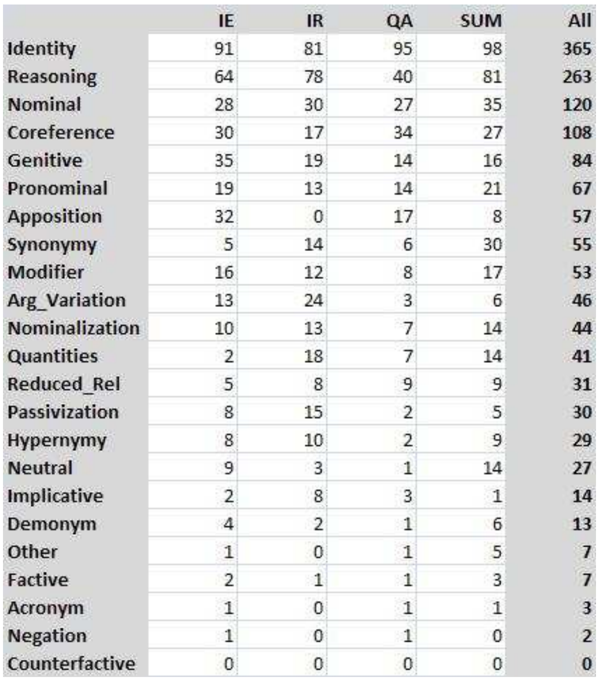
Table 4.5 shows the distribution of all twenty-three features[1] in the different subtasks, as well as in the complete positive entailment set. The features are listed in decreasing order of frequency in the complete annotated dataset. Figure 4.6 presents the corresponding chart.

We observe that the most frequent feature by far is Identity, which is hardly surprising, given the special nature of the feature. Overlooking Identity, we observe a long-tail statistical distribution in the data: A small high-frequency population is followed by a low-frequency population which gradually tails off, making up the majority of the graph.

The most frequent entailment feature is Reasoning, appearing altogether in 263 (65.75%) of the 400 annotated pairs. This indicates that a significant portion of the data involves deeper inferences; nonetheless the portion of the data which does not (137 out of 400 pairs; i.e. 34.25%) is considerable.

[1] We use the following abbreviations:

Arg_Variation	Argument Variation
Reduced_Rel	Reduced Relative



	IE	IR	QA	SUM	All
Identity	91	81	95	98	365
Reasoning	64	78	40	81	263
Nominal	28	30	27	35	120
Coreference	30	17	34	27	108
Genitive	35	19	14	16	84
Pronominal	19	13	14	21	67
Apposition	32	0	17	8	57
Synonymy	5	14	6	30	55
Modifier	16	12	8	17	53
Arg_Variation	13	24	3	6	46
Nominalization	10	13	7	14	44
Quantities	2	18	7	14	41
Reduced_Rel	5	8	9	9	31
Passivization	8	15	2	5	30
Hypernymy	8	10	2	9	29
Neutral	9	3	1	14	27
Implicative	2	8	3	1	14
Demonym	4	2	1	6	13
Other	1	0	1	5	7
Factive	2	1	1	3	7
Acronym	1	0	1	1	3
Negation	1	0	1	0	2
Counterfactive	0	0	0	0	0

Table 4.5: The distribution of individual entailment features in the positive entailment subsets of the RTE-2 test set.

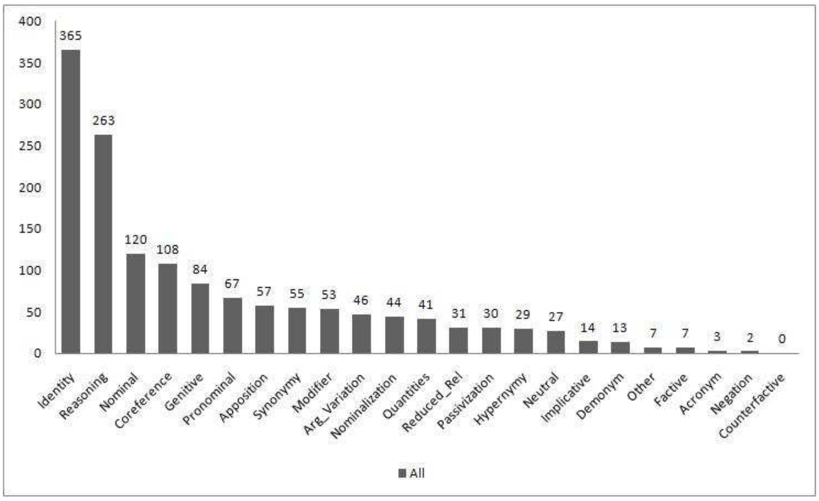


Figure 4.6: The chart of individual entailment features distribution in the complete positive entailment subset of the RTE-2 test dataset.

Finally there is a large number of features, including Negation, Acronym, Factive etc., that appear only marginally in the dataset.

The distribution of individual entailment feature occurrences differs across application settings. Figures 4.7, 4.8, 4.9 and 4.10 present the distributions separately for each one of the tasks IE, IR, QA and SUM, respectively.

It is obvious that certain entailment types are more common for some tasks than other. For instance, whereas 81% of the pairs in SUM involve Reasoning, only 40% of the QA pairs do so. Coreference also appears in IR half the times it does in QA (17% and 34%, respectively), while Genitive appears in 35% of the IE pairs but only in 14% of the QA pairs.

Further striking differences involve the features of Apposition and Synonymy. The former is the third most frequent feature of IE, occurring in 32% of the pairs, while at the same time it counts no occurrence in IR. The latter, Synonymy, demonstrates relatively low frequency in most of the tasks (5%, 14% and 6% for IE, IR and QA, respectively) but reaches a 30% frequency in SUM, constituting it the third most common entailment type of the task. Similar observations can be made for several other features.

Combinations of Entailment Features

Another parameter for the classification of the pairs is the number of different features they have been annotated for. As each alignment may carry one or more labels corresponding to different features, each pair, which is normally associated with several alignments, has a unique number of different annotated features.

Table 4.6 presents the distribution of the number of different features of the pairs individually in each of the four subtasks, as well as collectively in all. The non-informative feature Identity has been ignored for this purpose.

Clearly, the vast majority of the pairs is rather poor in entailment features. The most frequent type of pairs is the one with only 2 different features annotated, comprising 28% of the overall dataset. More than half (52%) of the total number of pairs have maximally 2 different features annotated and 75.2% have maximally 3 different features annotated. The pairs with richer annotation of 4 to 9 different features cover only the remaining 24.8% of the data.

Figure 4.11 shows the charts separately for each task. It is evident that, while the general trend towards poor annotation described in the previous paragraph remains present in all tasks, there are nevertheless certain differences among them. QA appears to be the task with the poorest annotation,

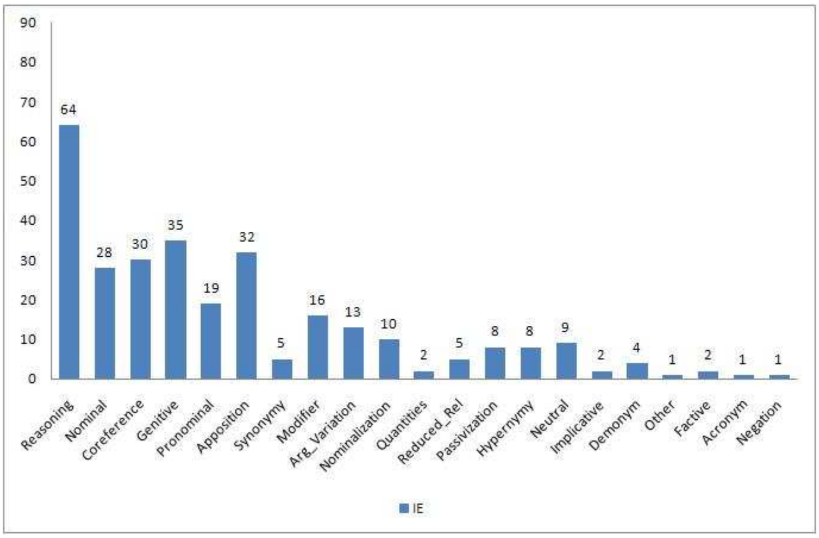


Figure 4.7: The chart of individual entailment features distribution in the positive entailment subset of the RTE-2 test dataset that corresponds to the IE task.

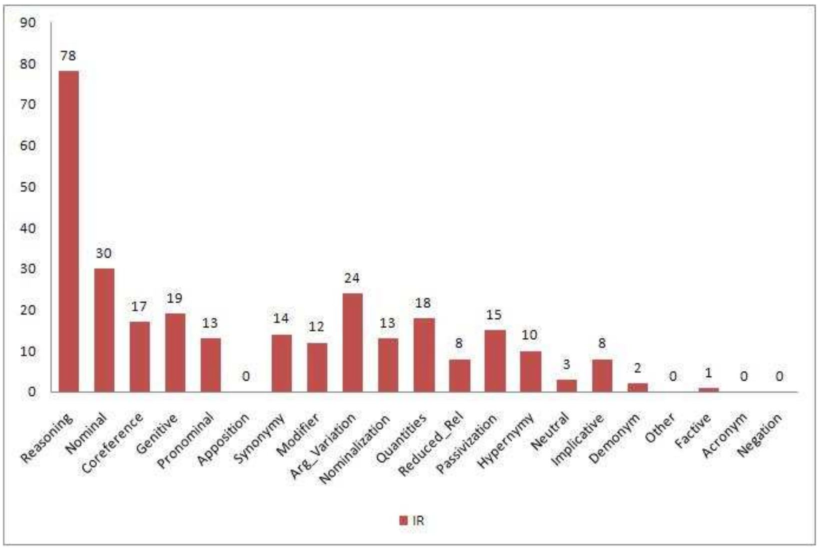


Figure 4.8: The chart of individual entailment features distribution in the positive entailment subset of the RTE-2 test dataset that corresponds to the IR task.

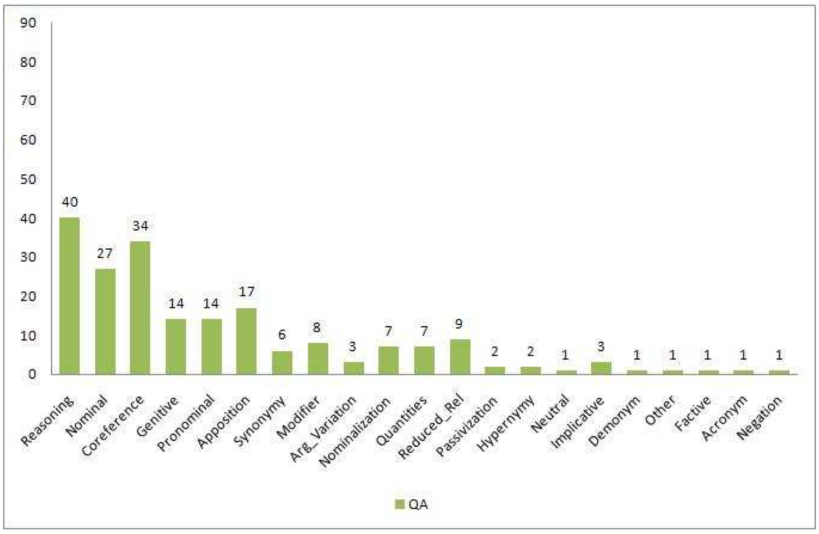


Figure 4.9: The chart of individual entailment features distribution in the positive entailment subset of the RTE-2 test dataset that corresponds to the QA task.

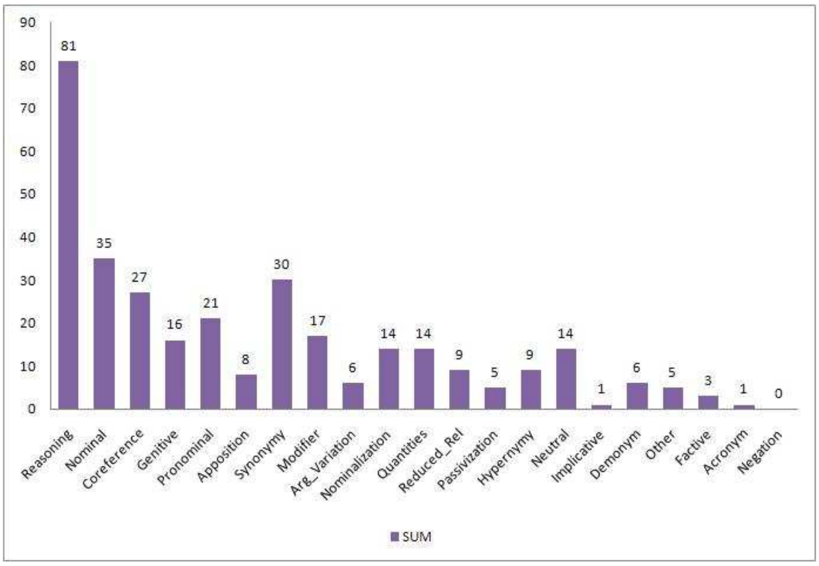
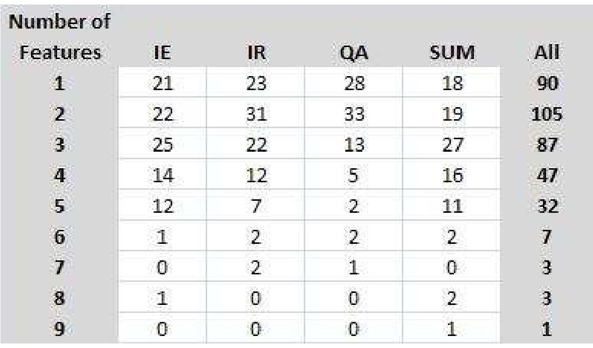


Figure 4.10: The chart of individual entailment features distribution in the positive entailment subset of the RTE-2 test dataset that corresponds to the SUM task.



Number of Features	IE	IR	QA	SUM	All
1	21	23	28	18	90
2	22	31	33	19	105
3	25	22	13	27	87
4	14	12	5	16	47
5	12	7	2	11	32
6	1	2	2	2	7
7	0	2	1	0	3
8	1	0	0	2	3
9	0	0	0	1	1

Table 4.6: The distribution of the number of different annotated features that correspond to the pairs of the positive entailment test set of RTE-2.

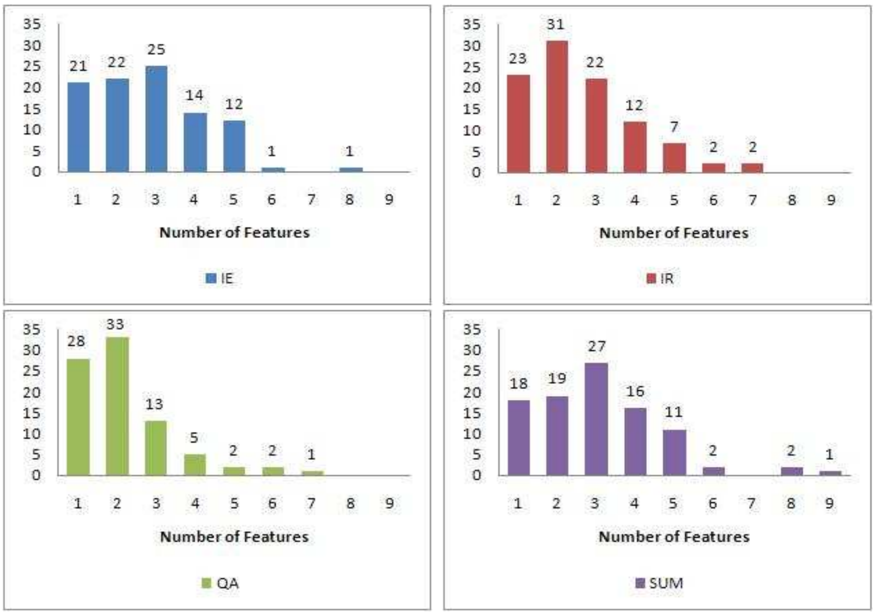


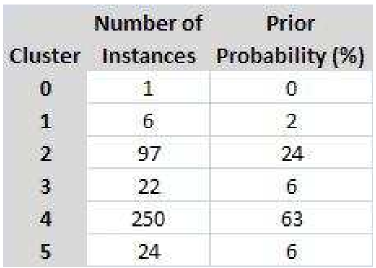
Figure 4.11: The charts of the distribution of the numbers of different entailment features annotated in the pairs of each of the four tasks.

having only approx 12% of its pairs annotated for more than 3 features. On the contrary, SUM exhibits a significantly higher richness of annotation, with one third of its pairs annotated for 4 to 9 different features.

However, factors such as the word length of the pairs must of course be taken into account for the interpretation of these results.

Data Clustering

To find out whether we can identify interesting patterns of similarity in the data, we conducted a clustering analysis experiment. For this purpose we used the Expectation-Maximization (EM) algorithm of the Weka clustering package (Witten and Frank, 2005). The result produced six clusters, as summarized in Table 4.7.



Cluster	Number of Instances	Prior Probability (%)
0	1	0
1	6	2
2	97	24
3	22	6
4	250	63
5	24	6

Table 4.7: The clusterer output. The number of clusters selected by cross validation is 6; the log likelihood score is -5.98983.

A single cluster, Cluster 4, seems to be covering the majority of the pairs. From the clusterer output it is found that the features mostly active in this cluster are Identity and Reasoning. Interesting is also to note that Cluster 0, with only one instance, consists of the pair 94 of the RTE-2 test set, which has the special property of having been annotated with 9 different features. As revealed in Table 4.6, this pair uniquely occupies the position of the most richly annotated pair in the dataset.

Nonetheless the results appear hard to interpret in useful ways, which is, after all, a typical problem related to clustering. Therefore we carry on the analysis of the data from other perspectives.

Entailment Types and Their Distribution

Apart from investigating the distributions of individual entailment features, we can also look into particular combinations of the feature occurrences in the datasets, which induce distinct types of entailment. One meaningful way of forming such types, compatible with the traditional distinctions among levels of entailment (e.g. Vanderwende et al. (2005); Bar-Haim et al. (2005); Clark et al. (2007)), is presented in Table 4.8.

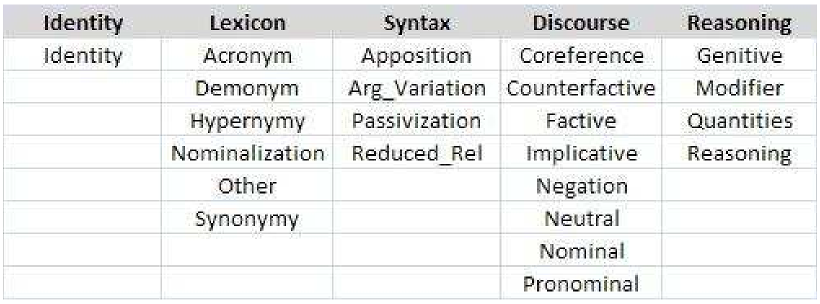
The entailment type **Identity** represents the pairs with all alignments labeled exclusively with the Identity feature, and no other kind of annotation. **Identity** is thus the simplest type of entailment. Note that, in general, Identity alignments are taken into account in this classification only in case they correspond to the sole kind of annotation of the pair, constituting the **Identity** entailment type. In any other case, where additional annotation is available, Identity alignments are ignored.

The entailment type **Lexicon** (**Lex**) represents the pairs with exclusively Alignment level annotation indicating Morphological (Acronym, Demonym, Nominalization, Other) and/or Ontological (Hypernymy, Synonymy) relations. **Syntax** (**Syn**) in turn stands for the pairs that implicate exclusively Arg_Variation alignments (including the special case of Passivization), while H may contain Supplemental expressions (Apposition, Reduced_Rel).

The entailment type **Discourse** (**Dis**), on the other hand, contains classical discourse features related to anaphora, as well as factivity features, whose resolution, involving presuppositions and implicatures, typically also refers back to properties of the discourse. It thus encompasses the pairs with annotation involving exclusively Coreference alignments, possibly accompanied by Anaphoric (Nominal, Pronominal) links at the Coreference level, while H may contain additional Context level annotation: Factivity (Counterfactive, Factive, Implicative, Neutral) or Negation.

Finally, the entailment type **Reasoning** (**Reas**) is representative of the pairs with exclusively Alignment level annotation involving one or more of the deeper inference features Genitive, Modifier, Quantities and Reasoning.

Table 4.9 shows the distribution of the different entailment types defined in this way in the datasets, listed in decreasing order of frequency for the



Identity	Lexicon	Syntax	Discourse	Reasoning
Identity	Acronym	Apposition	Coreference	Genitive
	Demonym	Arg_Variation	Counterfactive	Modifier
	Hypernymy	Passivization	Factive	Quantities
	Nominalization	Reduced_Rel	Implicative	Reasoning
	Other		Negation	
	Synonymy		Neutral	
			Nominal	
			Pronominal	

Table 4.8: The classification of the 23 features of the ARTE scheme into 5 entailment types.

overall positive entailment test set. The entailment types here correspond to the pairs whose entailment value can be judged *exclusively* by inferences related to the given types.[2] Moreover, Figure 4.12 displays the distribution for the overall set, while Figures 4.13, 4.14, 4.15 and 4.16 display separately the distributions for the individual subtasks.

It is remarkable that a considerable number of pairs (25 out of the 400 pairs; i.e. 6.25%) involves solely the **Identity** entailment type. However a significant portion of the pairs deals with deeper **Reasoning** entailments: 24.75% (99 out of 400 pairs) involve exclusively the **Reasoning** entailment type, while more than half of the pairs involve combinations of entailment types that include **Reasoning**.

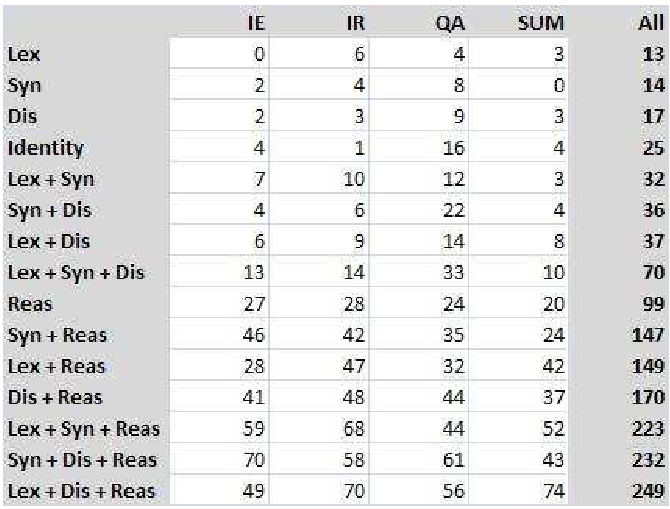
Obviously, the distributions vary to a large extent across the different tasks. For instance, IE contains no pairs that can be determined solely with **Lexicon**, whereas SUM contains no pairs that can be determined solely with **Syntax**. QA is the task with the most **Identity** entailments (64% of all **Identity** entailments), while IR is the one with the fewest (only 4%). In parallel, QA contains the majority of entailments that can be determined solely by **Syntax** and **Discourse** (**Syn** + **Dis**), holding approx 61% of the **Syn** + **Dis** pairs, while all the other tasks are almost equally poorer in this particular entailment combination.

Finally, Table 4.10 presents the distribution of entailment types *inclusively*, i.e., the frequencies of pairs that can be determined with use of the type in question, possibly combined with other types. Figure 4.17 shows the corresponding charts for the different tasks of RTE-2.

Reasoning appears to be the most frequent type of entailment here by far, both in the overall dataset and in the subsets belonging to the individual tasks. However only 51% of the pairs in QA involve **Reasoning**, while in any other task, **Reasoning** entailments occupy at least an 83% portion.

As a further observation, the proportion of **Lexicon** to **Syntax** entailments approximates 1:2 for the IE and QA, but 2:1 for the SUM task, which is also the one with the lowest proportion of **Syntax** entailments. **Discourse** entailments appear to be in a roughly equal distribution across the tasks.

[2] For example, **Lex** + **Syn** involves the pairs that are annotated with features belonging to one or both of the types **Lex** and **Syn**, and no additional features. The type **Dis** + **Lex** + **Reas** + **Syn** is not listed, since it uninterestingly encompasses all the pairs that have any annotation other than that of Identity.



	IE	IR	QA	SUM	All
Lex	0	6	4	3	**13**
Syn	2	4	8	0	**14**
Dis	2	3	9	3	**17**
Identity	4	1	16	4	**25**
Lex + Syn	7	10	12	3	**32**
Syn + Dis	4	6	22	4	**36**
Lex + Dis	6	9	14	8	**37**
Lex + Syn + Dis	13	14	33	10	**70**
Reas	27	28	24	20	**99**
Syn + Reas	46	42	35	24	**147**
Lex + Reas	28	47	32	42	**149**
Dis + Reas	41	48	44	37	**170**
Lex + Syn + Reas	59	68	44	52	**223**
Syn + Dis + Reas	70	58	61	43	**232**
Lex + Dis + Reas	49	70	56	74	**249**

Table 4.9: The distribution of the different entailment types, as defined in Table 4.8. The frequencies correspond to the number of pairs that can be determined *exclusively* with use of the entailment types in question.

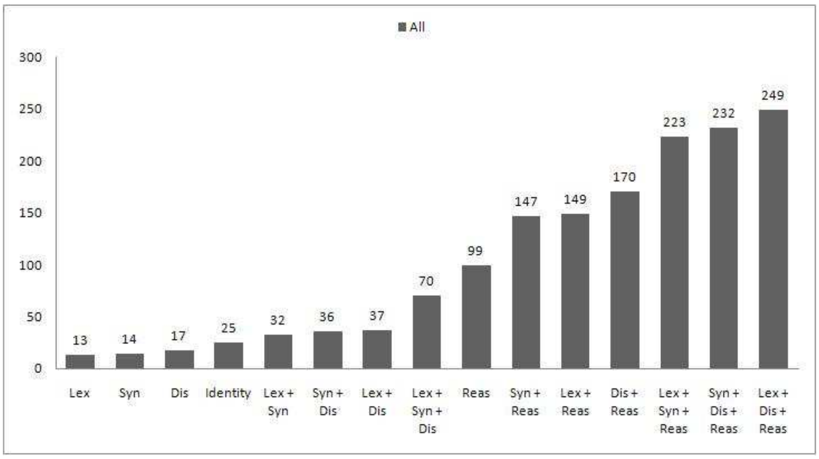


Figure 4.12: The chart of the distribution of different entailment types in the complete positive entailment subset of the RTE-2 test dataset, drawn from Table 4.9.

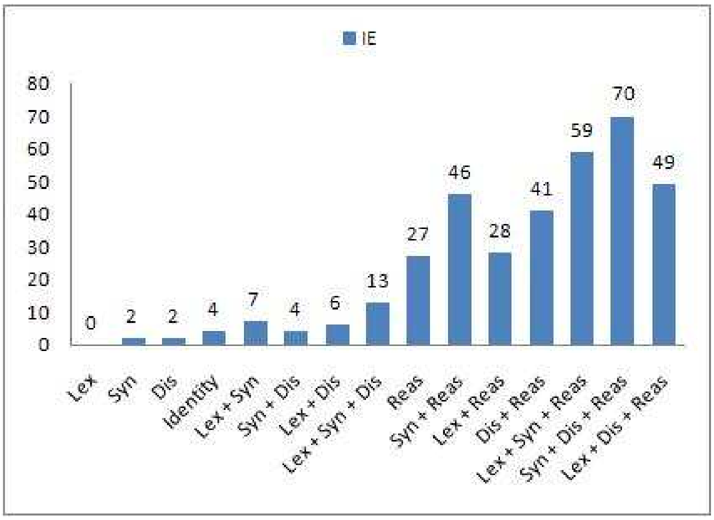


Figure 4.13: The chart of the distribution of different entailment types in the IE task, drawn from Table 4.9.

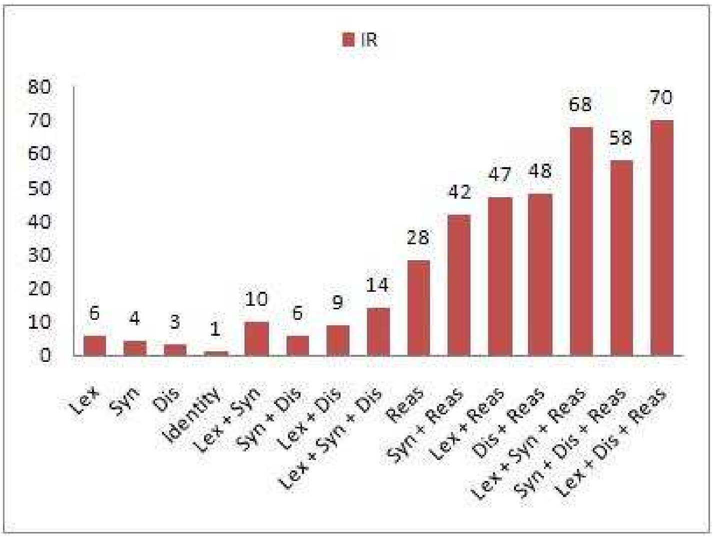


Figure 4.14: The chart of the distribution of different entailment types in the IR task, drawn from Table 4.9.

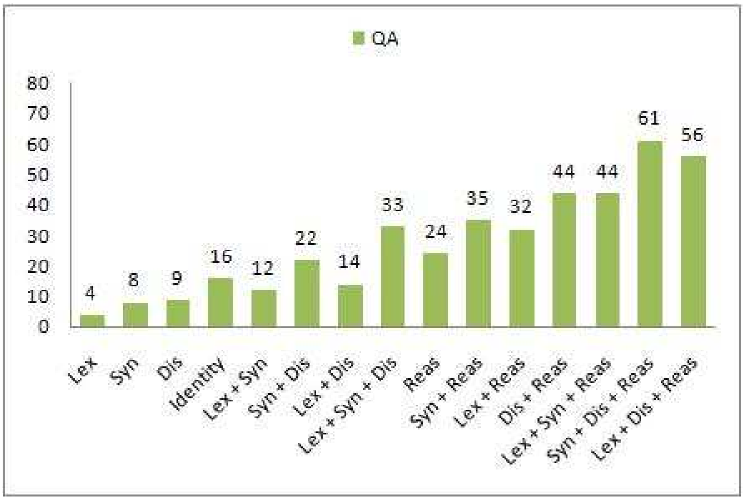


Figure 4.15: The chart of the distribution of different entailment types in the QA task, drawn from Table 4.9.

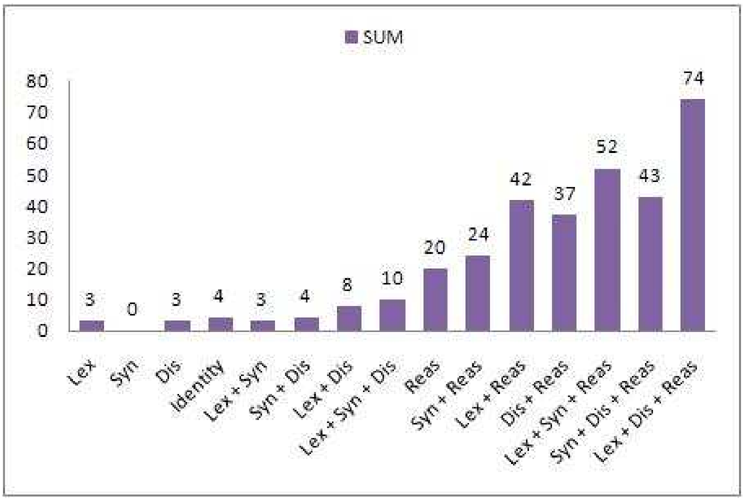
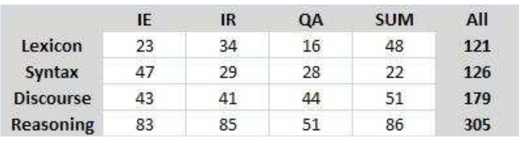


Figure 4.16: The chart of the distribution of different entailment types in the SUM task, drawn from Table 4.9.



	IE	IR	QA	SUM	All
Lexicon	23	34	16	48	121
Syntax	47	29	28	22	126
Discourse	43	41	44	51	179
Reasoning	83	85	51	86	305

Table 4.10: The distribution of the different entailment types, *inclusively*: In contrast to Table 4.9, the frequencies here correspond to the number of pairs that can be determined with use of the entailment type in question, possibly in combination with other types.

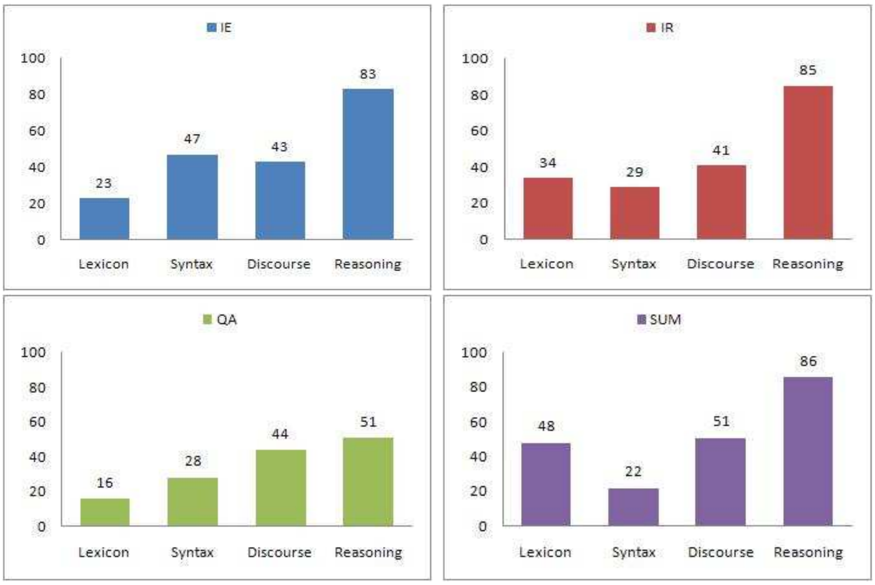


Figure 4.17: The charts of the distributions of Table 4.10 separately for each task.

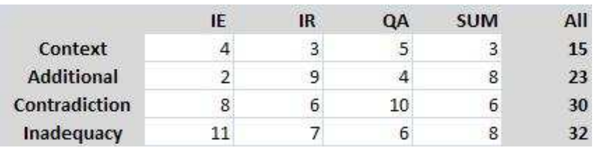
4.2.2 The Negative Entailment Dataset

As presented in Section 3.2, our experimental negative entailment scheme comprises only four features: Context, Additional, Contradiction and Inadequacy. Their distribution in the 100 pairs of the test set we annotated is shown in Table 4.11 in decreasing order of frequency in the sample.

Clearly, Inadequacy and Contradiction are the most common types of non-entailment, while Context is the most scarce. A remarkable result is that the proportion of contradictions found in the annotated negative entailment dataset (30%) is in total agreement with the one reported by Manning et al. (2007), regarding contradictions in RTE challenges prior to the third.

Figure 4.18 presents the results for each task individually, indicating differences among the tasks also in the case of negative entailment. For instance, IR appears to be considerably richer than IE in non-entailments due to the presence of additional information in H (36% vs. 8%, respectively).

Given the sparsity of the data in the sample, nonetheless, conclusions must in this case be drawn with particular cautiousness.



	IE	IR	QA	SUM	All
Context	4	3	5	3	15
Additional	2	9	4	8	23
Contradiction	8	6	10	6	30
Inadequacy	11	7	6	8	32

Table 4.11: The distribution of non-entailment types in a random sample of 100 negative entailment pairs of the RTE-2 test set, equally distributed among the four tasks.

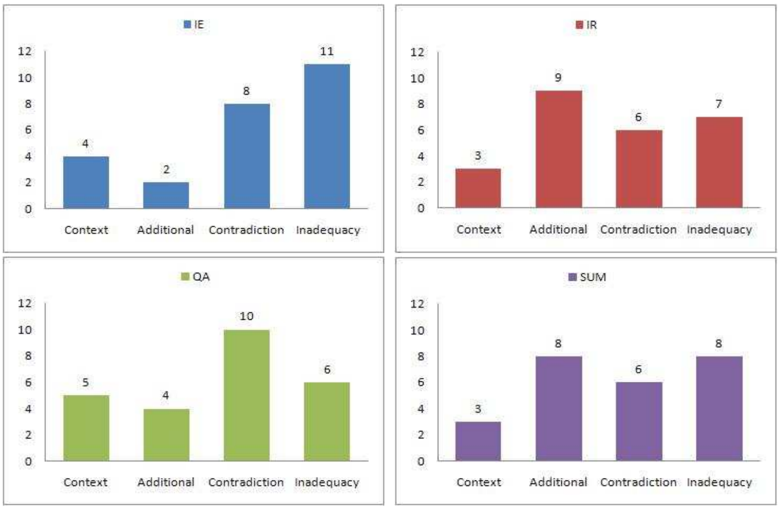


Figure 4.18: The charts of the distributions of Table 4.11 separately for each task.

4.2.3 Discussion

The annotation of the RTE-2 dataset applying the ARTE scheme has enabled new ways of evaluating textual entailment corpora, which could not be available otherwise. However, the reliability and significance of the results stemming from this evaluation are definitely subject to the annotation process itself, and, most importantly, to the annotation scheme and guidelines.

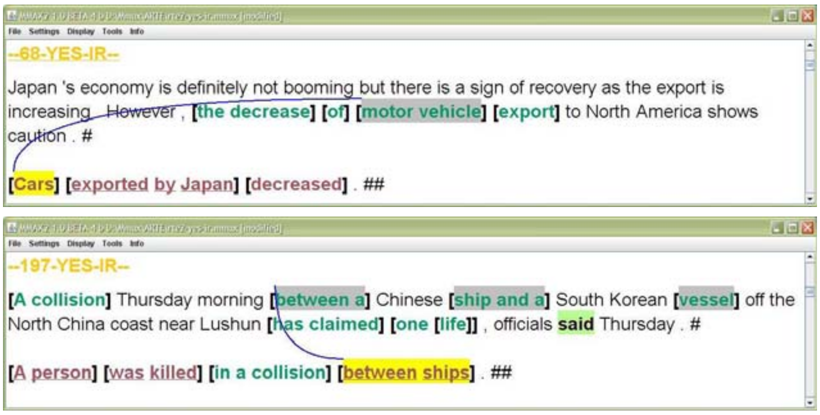
One specific problem arising from the annotation scheme applied is that, for practical reasons, the twenty-three features employed are not equally fine-grained. In particular, the feature Reasoning uniformly embraces a wide range of dissimilar forms of textual entailment inferences (see end of Subsection 3.1.1), whereas most of the other features are rather more specific and of limited scope.

As an outcome, the frequency of Reasoning in the data reaches such rates that its comparability to the rest of the features is arguably put to question. This characteristic naturally extends to the entailment type of **Reasoning**, which builds on the homonymous feature, and all its combinations.

The interpretation of the statistical results of the present study must therefore take this fact into account. On the other hand, given the novel and fairly experimental status of the effort, as well as its apparent limitations, we believe that the analysis conducted has made a contribution of several interesting remarks in the research field of textual entailment recognition.

Finally, to offer another point of view to the discussion about the usability of the annotation, we would like to call attention to a few indicative pairs whose entailment judgments have been considered as rather problematical by the annotators.

In the pairs of Figure 4.19, which are both associated with positive entailment values according to the gold standard, the entailment relation does not seem obvious. On the one hand, there is a clear hypernymy relation[3] among the aligned expressions (a *car* is a kind of *motor vehicle*; a *ship* is a kind of *vessel*).


File Settings Display Tools Info
File Settings Display Tools Info

--68-YES-IR--

Japan 's economy is definitely not booming but there is a sign of recovery as the export is increasing . However , [the decrease] [of] [motor vehicle] [export] to North America shows caution . #

[Cars] [exported by Japan] [decreased] . ##

--197-YES-IR--

[A collision] Thursday morning [between a] Chinese [ship and a] South Korean [vessel] off the North China coast near Lushun [has claimed] [one [life]] , officials said Thursday . #

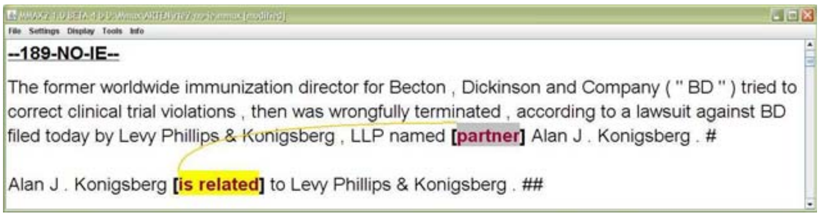
[A person] [was killed] [in a collision] [between ships] . ##

Figure 4.19: Examples of disputed entailment involving the direction of hypernymy relations.

[3] Retrieved from the WordNet lexical database: http://wordnet.princeton.edu/perl/{webwn?s=car, webwn?s=ship}.

On the other hand, the direction of the hypernymy relation in these cases does not justify an entailment relation, since in both examples the general concept appears in H and the more specific in T. A positive entailment would require the opposite order.

Figure 4.20, lastly, presents an example of non-entailment, as indicated in the gold standard, which nonetheless the annotators have had difficulty in justifying.



--189-NO-IE--

The former worldwide immunization director for Becton , Dickinson and Company (" BD ") tried to correct clinical trial violations , then was wrongfully terminated , according to a lawsuit against BD filed today by Levy Phillips & Konigsberg , LLP named [partner] Alan J . Konigsberg . #

Alan J . Konigsberg [is related] to Levy Phillips & Konigsberg . ##

Figure 4.20: This pair, which has officially been assigned a negative entailment value, raises questions regarding the appropriateness of the gold standard.

Such controversial pairs were not met too frequently in the annotated datasets. However, it should be pointed out that the problem of producing an annotation of high quality and standards, which can be beneficial to the evaluation of the data, is strongly intertwined with the problem of creating a reliable gold standard annotation of entailment, able to function as a basis for further work.

4.3 Summary

The analysis of the datasets with respect to shallow and deep measures conducted in the previous sections has set the conditions for a wide range of direct observations regarding the less than perfectly clear nature of RTE.

The results indicate that a large number of semantically and linguistically challenging entailment types is generally only marginally present in the data, while very few types occur with high frequency (see Figure 4.6). Moreover, as the overlap–entailment correlation analysis of Subsection 4.1.3 suggests, the datasets are compiled in a way rather favorable to shallow bag-of-words approaches.

From another standpoint, the four individual tasks that partition the set have been shown to be substantially different in several aspects, since, among other differences, entailment types are distributed unequally among them.

Such differences, as well as differences among positive and negative entailment pairs, are expected to significantly improve systems' performance, if taken into account.

Chapter 5

Analysis of the RTE-2 Systems' Performance

The previous chapter attempted an evaluation of the RTE-2 dataset per se, i.e. by means of investigating its shallow and deep linguistic properties, and irrespective of real systems' performance on it.

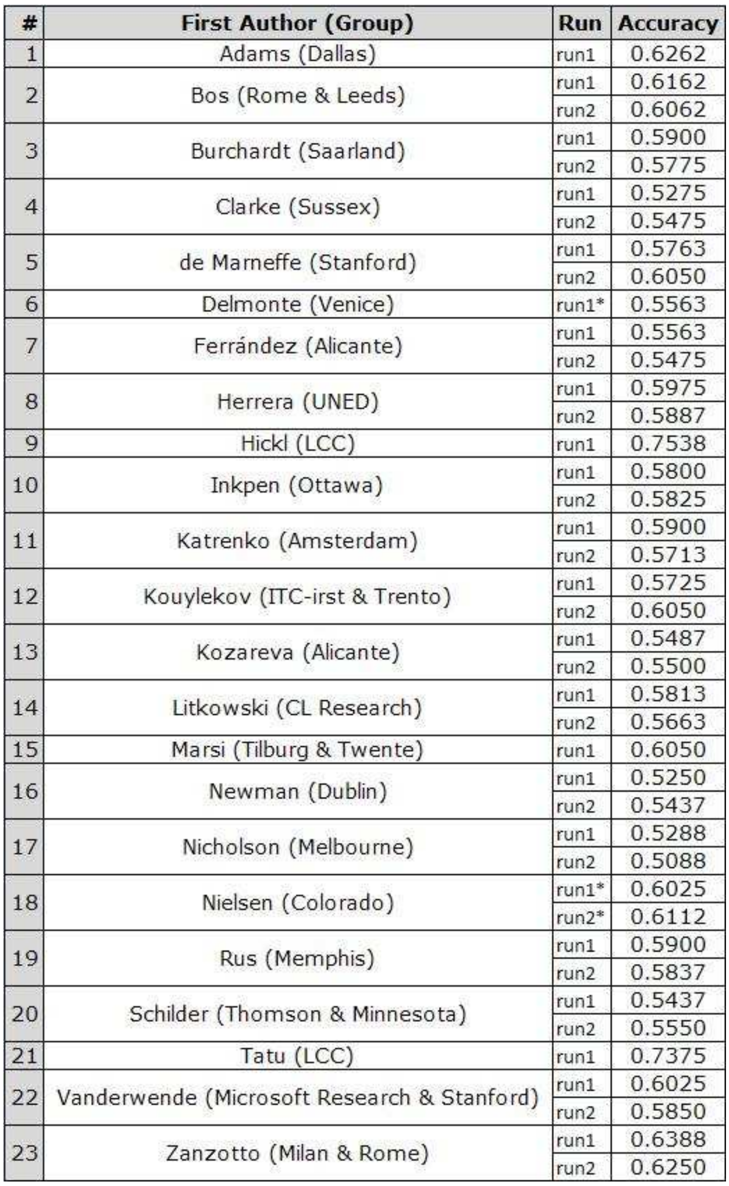
This investigation makes room for posing several further intriguing questions: How do different types of textual entailment systems perform on different subsets of the dataset? What can be discovered about the correlations between systems' performance and characteristics of the datasets, on the one hand, and systems' performance and internal properties of their architecture, on the other?

The present chapter addresses these issues. The RTE-2 dataset is studied here in the light of systems' performance on it both collectively, and by discriminating against different textual entailment types, as well as different system types of RTE.

5.1 Collective Systems' Performance

As reported in (Bar-Haim et al., 2006), 23 teams participated in the Second RTE Challenge, each being allowed to submit results of up to 2 systems. Since many of the participants chose to make use of this option, and provided results of 2 runs, the total number of different systems competing was as high as 41.

As mentioned earlier, these results have been evaluated with two criteria: accuracy and, optionally, average precision, as a measure for the ranking of pairs according to their entailment confidence (when applicable). Table 5.1 summarizes the submission results for the evaluation criterion of accuracy,



#	First Author (Group)	Run	Accuracy
1	Adams (Dallas)	run1	0.6262
2	Bos (Rome & Leeds)	run1	0.6162
		run2	0.6062
3	Burchardt (Saarland)	run1	0.5900
		run2	0.5775
4	Clarke (Sussex)	run1	0.5275
		run2	0.5475
5	de Marneffe (Stanford)	run1	0.5763
		run2	0.6050
6	Delmonte (Venice)	run1*	0.5563
7	Ferrández (Alicante)	run1	0.5563
		run2	0.5475
8	Herrera (UNED)	run1	0.5975
		run2	0.5887
9	Hickl (LCC)	run1	0.7538
10	Inkpen (Ottawa)	run1	0.5800
		run2	0.5825
11	Katrenko (Amsterdam)	run1	0.5900
		run2	0.5713
12	Kouylekov (ITC-irst & Trento)	run1	0.5725
		run2	0.6050
13	Kozareva (Alicante)	run1	0.5487
		run2	0.5500
14	Litkowski (CL Research)	run1	0.5813
		run2	0.5663
15	Marsi (Tilburg & Twente)	run1	0.6050
16	Newman (Dublin)	run1	0.5250
		run2	0.5437
17	Nicholson (Melbourne)	run1	0.5288
		run2	0.5088
18	Nielsen (Colorado)	run1*	0.6025
		run2*	0.6112
19	Rus (Memphis)	run1	0.5900
		run2	0.5837
20	Schilder (Thomson & Minnesota)	run1	0.5437
		run2	0.5550
21	Tatu (LCC)	run1	0.7375
22	Vanderwende (Microsoft Research & Stanford)	run1	0.6025
		run2	0.5850
23	Zanzotto (Milan & Rome)	run1	0.6388
		run2	0.6250

Table 5.1: Systems accuracy results of RTE-2. Runs marked with * indicate resubmission after publication of the official results, allowed only in case of a bug fix.

which, representing the percentage of pairs correctly judged, constitutes the main evaluation measure.

Note that the performance figures presented in this and the following sections involve exclusively the test set, and not the development set of RTE-2, since this is the dataset on which system results are evaluated.

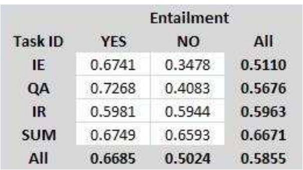
5.1.1 Collective Performance Across Tasks and Entailment Values

From the individual results[1] of the forty-one systems of Table 5.1 on a pair of the RTE-2 test set we compute the average system performance on the given pair. This is defined as the ratio of correct system answers over the total number of system answers:

$$\text{Performance} = \frac{\#\ \text{Correct System Answers}}{\#\ \text{System Answers}} \tag{5.1}$$

The average system performance on the RTE-2 sets of pairs, listed in decreasing order of overall performance, is presented in Table 5.2. Figure 5.1 displays the distribution of average system performance on the different tasks of RTE-2.

In agreement to the participants' reports, the figures show that the four application settings are not balanced with respect to systems' performance.



	Entailment		
Task ID	YES	NO	All
IE	0.6741	0.3478	**0.5110**
QA	0.7268	0.4083	**0.5676**
IR	0.5981	0.5944	**0.5963**
SUM	0.6749	0.6593	**0.6671**
All	**0.6685**	**0.5024**	**0.5855**

Table 5.2: The distribution of average system performance on the RTE-2 datasets, as defined by (5.1). The complete test set of 800 pairs is considered here.

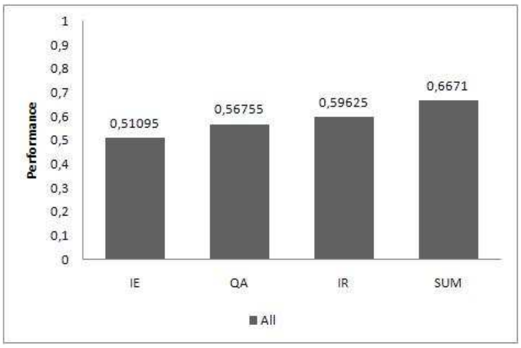


Figure 5.1: The chart of average system performance on the four tasks of RTE-2.

[1] Publicly available at http://www.pascal-network.org/Challenges/RTE2/Results.

The most challenging task appears to be IE, with average performance only slightly higher than 0.50. QA and IR follow, and finally, SUM rightfully claims the title of the "easiest" task, with a performance of approx 0.67, significantly higher than that of any other task.

Figure 5.2, which displays the distribution of average system performance on the eight different subcorpora, allowing for a differentiation between the positive and negative entailment pairs, presents a particularly interesting picture. On the two tasks with the lowest performance, IE and QA, we observe a great difference between their positive and negative entailment subsets, while on the two "easier" ones, IR and SUM, this difference is much more moderate.

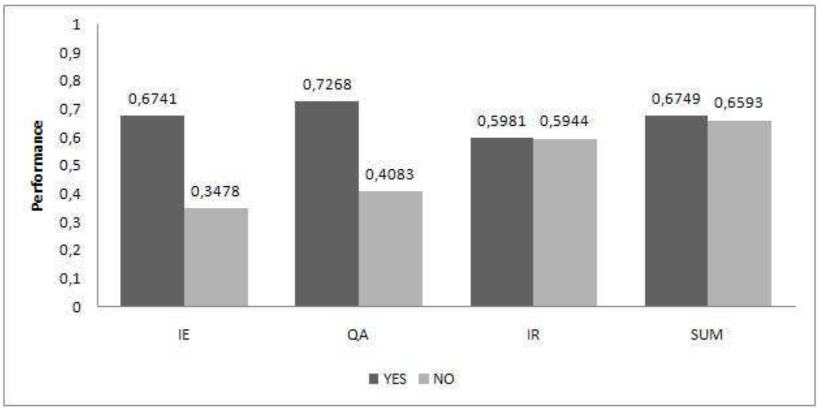


Figure 5.2: The chart of average system performance on the eight different subcorpora of RTE-2.

5.1.2 Overlap–Entailment Correlation and Performance

The picture of systems' performance in the datasets described in the previous subsection naturally brings to memory the overlap–entailment correlation statistics of Subsection 4.1.3[2].

Indeed, comparing Figures 4.5 and 5.2, which display the distributions of overlap–entailment correlation and system performance on the datasets, respectively, we discover that the two graphs follow an identical pattern. Furthermore, the similarity extends to the case of the complete sets corresponding to the individual tasks, as presented by Figures 4.4 and 5.1.

[2] As a matter of fact, the sample analyzed in Subsection 4.1.3 is the complete RTE-2 dataset, including both development and test set. This is not the case here, where only the test set is analyzed. However, given that the split is arbitrary, we do not suppose that the general trends observed will differ.

Hence it seems that there is possibly an underlying connection between the two notions of overlap–entailment correlation and system performance.

More concretely, whereas some of the semantic-linguistic aspects of entailment we examined in Section 4.2, such as the combinations of entailment features annotated, are not able to justify the system performance results, the concept of overlap–entailment correlation appears to be a good candidate for predicting systems' performance.

For instance, among the positive entailment subsets, the one corresponding to SUM was found to have received the richest annotation, with the highest number of different entailment features annotated (see Table 4.6), as well as the greatest portion of the demanding **Reasoning** type annotation (see Table 4.10). Considering this, one might expect SUM to be one of the hardest tasks for RTE-2. Nonetheless, as Table 5.2 indicates, it is rather the easiest.

Looking at Table 4.4, the reason for this may become clear: SUM is the task with the highest average overlap–entailment correlation, which remains relatively stable across entailment values. Since the majority of RTE-2 systems incorporate overlap-based techniques, it is natural that a dataset with high overlap–entailment correlation will also gain a good system performance result. Therefore, as foreseen in Subsection 4.1.3, the notion of overlap–entailment correlation seems to be particularly important in explaining and predicting the performance of real systems on the datasets.

This provides a good justification for the wide use of shallow bag-of-words approaches in RTE: They deal with the problem of textual entailment extraordinarily well with respect to the current evaluation setup. However, such mechanisms for judging textual entailment may not appeal to human intuitions. Even more importantly, they are prone to failure, as it is neither hard to generate cases of low overlap–entailment correlation, nor is there evidence of their scarcity in real-life scenarios.

On top of that, if the RTE Challenges are to promote textual entailment research, and create the conditions for "smarter" and more efficient textual entailment engines, then they need to pose substantial challenges to these engines. It is rather questionable whether textual entailment research can overcome its current limitations and truly advance, as long as the training and testing datasets guarantee high rates of success even to naive methods of low potential.

5.1.3 Collective Performance Across Entailment Types

In Subsection 5.1.1 we saw that system performance varies across the different task IDs and entailment values of the RTE-2 dataset. But how does the picture evolve if we consider different entailment types?

Linguistic intuitions may be the reason for suspecting that the systems do not handle all types of entailment with equal success. The entailment type of **Identity**, for instance, which imposes no great linguistic challenges to systems, is expected to reach higher rates of system performance than other, more complex types, such as **Reasoning**.

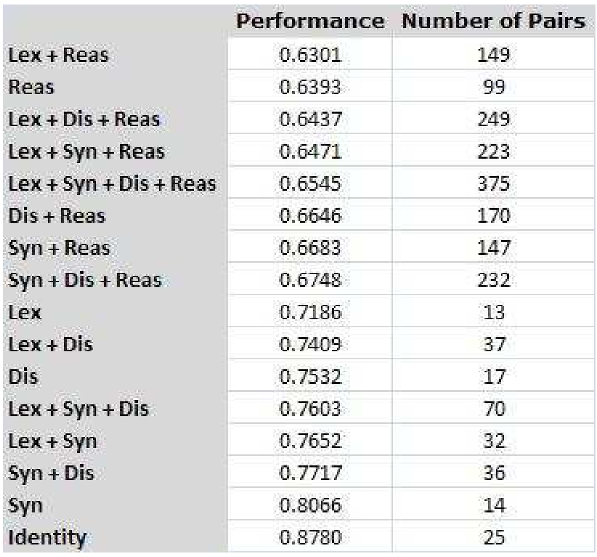
To find out to what extent such suspicions are verified by facts, and how precisely system performances relate to entailment types, we compute the system performance notion of Definition (5.1) for the datasets corresponding to different entailment types, as determined in Table 4.8 of Section 4.2.

The result is depicted in Table 5.3, in increasing order of system performance. The frequency of each entailment type, as presented in Table 4.9 of Section 4.2, is replicated here for convenience. Similarly to Table 4.9, the figures in this table correspond to the pairs of the positive entailment test set of RTE-2 whose entailment value can be determined *exclusively* with use of one or more of the given entailment types. Figure 5.3 presents the accompanying chart.

Clearly the initial predictions meet reality: The entailment type of **Identity** is indeed the "easiest" case for systems, claiming the highest performance achieved (approx 0.88). Other "easy" types of entailment include **Syntax** and all its combinations with **Lexicon** and **Discourse**, though with significantly lower performance values than **Identity**.

At the far left side of the graph, on the other hand, the entailment type of **Reasoning** and all its combinations with other types appear to inflict difficulties on systems and only achieve performances of approx 0.63 or slightly higher. The entailment types of **Lexicon**, **Discourse** and their combination occupy the middle area of the graph, corresponding to intermediate levels of system performance.

This picture is to a degree preserved, but not identical, when we examine the data from the non-exclusive perspective: Table 5.4 presents the distribution of system performance figures across entailment types *inclusively*, i.e. representing the pairs whose entailment value can be determined with use of the given entailment type, possibly in combination with other types of entailment. Again the frequencies of the types, as presented in the corresponding Table 4.10 of Section 4.2, are reproduced. Figure 5.4 shows the accompanying chart.



	Performance	Number of Pairs
Lex + Reas	0.6301	149
Reas	0.6393	99
Lex + Dis + Reas	0.6437	249
Lex + Syn + Reas	0.6471	223
Lex + Syn + Dis + Reas	0.6545	375
Dis + Reas	0.6646	170
Syn + Reas	0.6683	147
Syn + Dis + Reas	0.6748	232
Lex	0.7186	13
Lex + Dis	0.7409	37
Dis	0.7532	17
Lex + Syn + Dis	0.7603	70
Lex + Syn	0.7652	32
Syn + Dis	0.7717	36
Syn	0.8066	14
Identity	0.8780	25

Table 5.3: The distribution of system performance, according to the definition (5.1), across the different entailment types, as defined by Table 4.8. Entailment types are meant here in the *exclusive* sense. Their frequency distribution is copied from Table 4.9 for convenience.

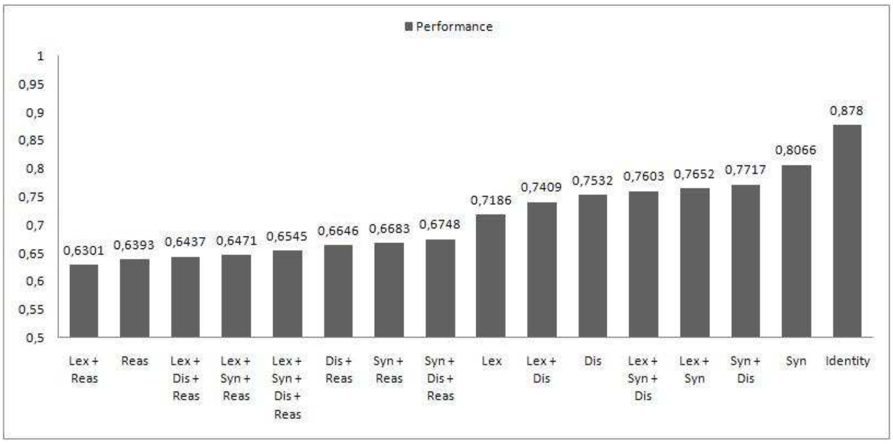
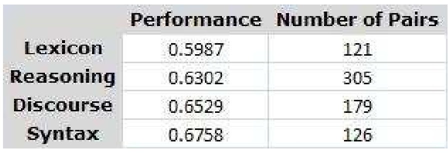


Figure 5.3: The chart of the distribution of system performance across the different entailment types, drawn from Table 5.3.



	Performance	Number of Pairs
Lexicon	0.5987	121
Reasoning	0.6302	305
Discourse	0.6529	179
Syntax	0.6758	126

Table 5.4: The distribution of system performance across different entailment types of the RTE-2 positive entailment test set. In contrast to Table 5.3, the entailment types here are meant *inclusively*. The frequency distribution of the entailment types is reduplicated from Table 4.10 for convenience.

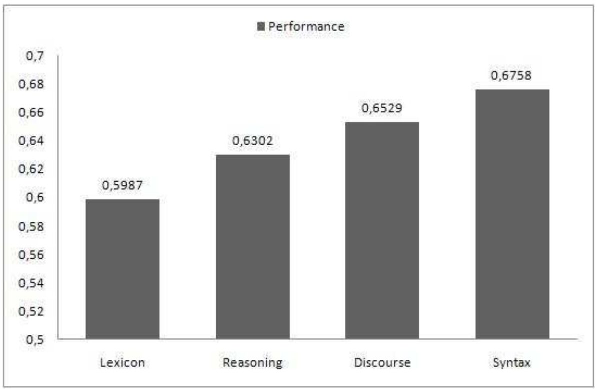


Figure 5.4: The chart of the distribution of system performance across different entailment types, as in Table 5.4.

Evidently the pairs that achieve the best performance are the ones that include entailments of the type of **Syntax**. **Discourse** pairs follow in the rank, followed in turn by **Reasoning** and then by **Lexicon** entailment pairs. This suggests that entailments that involve **Lexicon**, possibly together with other types, may be even harder for systems than entailments that involve **Reasoning**. The combination of the two, **Lex** + **Reas**, is after all, as derived by Table 5.3, the "hardest" type of entailment.

Finally, let us remark that the performance figures for all entailment types without exception are significantly higher than the 0.50 baseline, and even slightly higher than the approx 0.59 average overall system performance, reported in Table 5.2 of Subsection 5.1.1. Indeed, the lowest performance rate from the exclusive perspective—the one achieved for **Lex** + **Reas**—is as impressive as approx 0.63, while the lowest one from the inclusive perspective (approx 0.60 for **Lexicon**), though lower, is still higher than the overall average.

This is explained by the fact that the average system performance on the positive entailment subset of RTE-2 is significantly higher than the one corresponding to the negative: approx 0.69 as opposed to a baseline of approx 0.50, respectively (see Table 5.2). Therefore the general performance picture illustrated here, as well as in the following section, should not mistakenly be considered as representative of the complete dataset, as it describes solely the positive entailment subset of RTE-2.

5.2 System Types and Their Performance

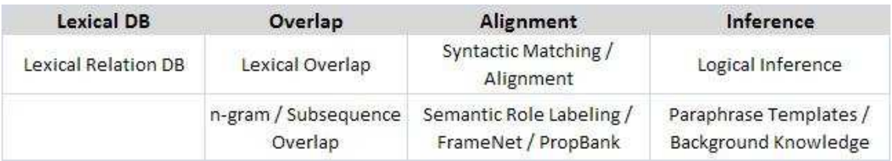
The fact that in RTE-2 different entailment types achieve different performance rates raises questions about the existence of types of systems that perform optimally for a certain entailment type, and perhaps sub-optimally for other types of entailment. Among the individual systems participating in the challenge—many of them independently built, taking dissimilar approaches to textual entailment recognition, implementing different techniques and utilizing different modules—, which groups of systems perform best for each different type of input?

An investigation in this direction could teach us much about ways to construct better entailment engines, optimally configured according to the nature of the data they are meant for. In the present section we take some first steps by analyzing the performance of four distinct system types of RTE-2 on the positive entailment subcorpora.

Bar-Haim et al. (2006) provide a useful overview of the forty-one systems participating in RTE-2 by classifying them into nine broad categories, according to their components. Based on this system description, we define four basic system types—**Lexical DB**, **Overlap**, **Alignment** and **Inference** systems—, as in Table 5.5.

Into the type of lexical relation database (**Lexical DB**) systems fall the ones that make use of lexical overlapping methods, based on lexicons such as WordNet (Fellbaum, 1998).

The type **Overlap** comprises the systems that function either merely by lexical overlap, or by using statistical techniques such as n-gram matching



Lexical DB	Overlap	Alignment	Inference
Lexical Relation DB	Lexical Overlap	Syntactic Matching / Alignment	Logical Inference
	n-gram / Subsequence Overlap	Semantic Role Labeling / FrameNet / PropBank	Paraphrase Templates / Background Knowledge

Table 5.5: The definition of system types according to their components, as derived from the system description in (Bar-Haim et al., 2006).

and subsequence overlapping. Note that the category of mere lexical overlap comprises only two systems, the ones for which no component is indicated in Table 2 of (Bar-Haim et al., 2006): run2 of Katrenko (Amsterdam) and run1 of Litkowski (CL Research).

The systems that use some kind of syntactic matching or alignment, e.g. relation matching and tree edit distance algorithms, or semantic role labeling method, e.g. semantic annotation induced from the FrameNet lexical resource (Baker et al., 1998), belong to the **Alignment** type. Finally, systems exploring logical inference (e.g. logic provers) or paraphrase templates and background knowledge approaches, including inference rules, constitute the **Inference** system type.

Table 5.6 presents the classification of the forty-one RTE-2 systems in the system types defined this way. Obviously the system types are not mutually exclusive; depending on its components description, a system can be classified into more than a single system type. Apart from that, the groups of systems formed by the defined types are not equally large: The vast majority of the systems fall into the **Lexical DB** or **Alignment** group—and often both—, but much fewer systems comprise the **Overlap** and **Inference** groups.

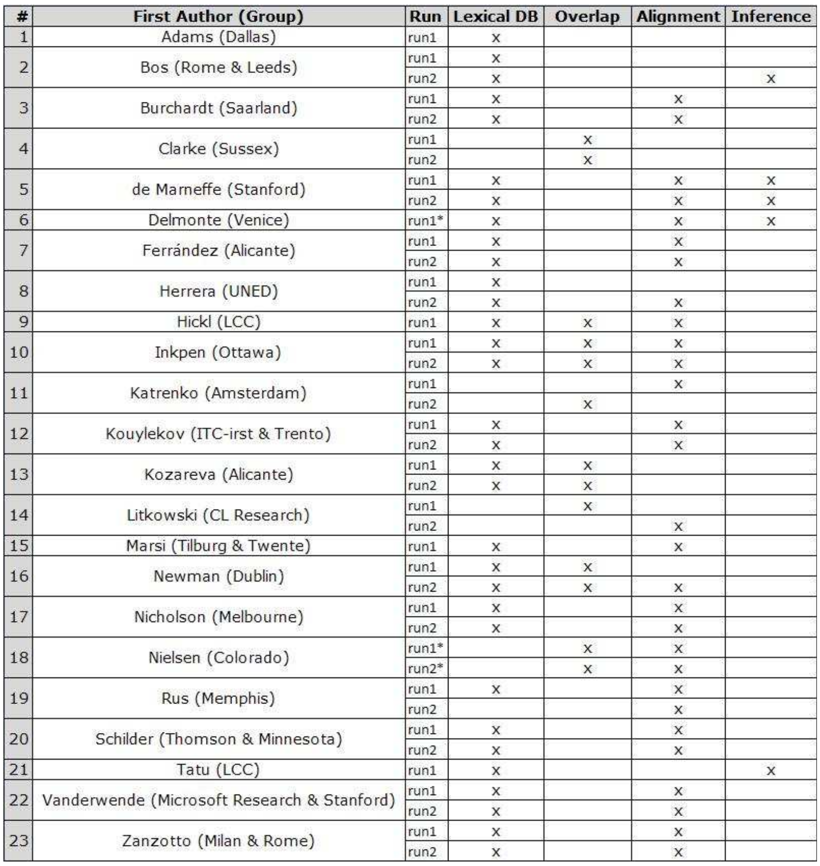
Definition (5.1) of the notion of performance on each pair can now be applied to the groups of systems formed according to their system type, as in Table 5.6. We analyze in this way the dataset corresponding to the positive entailment pairs of the RTE-2 test set, consisting of 400 pairs. Examined are both aspects of task IDs and types of entailment.

5.2.1 Performance Across Tasks

Table 5.7 presents the distribution of average system performance of each system type across the four tasks IE, IR, QA and SUM of RTE-2, while the distribution for the overall dataset is displayed in Figure 5.5.

Clearly the four system types perform differently on the datasets. **Inference** is the system type with the poorest performance (approx 0.63), outperformed even by **Overlap** (approx 0.65). **Lexical DB** and **Alignment** systems appear to achieve the best results in the datasets, with almost equal average performances of approx 0.68.

Figure 5.6 shows the charts of the individual distributions of performance of each system type across the four application settings. Again differences among the system types are evident. Although IR is the task on which all system types uniformly demonstrate their lowest performance, **Lexical DB** outperforms all other types on this task, achieving a performance signifi-



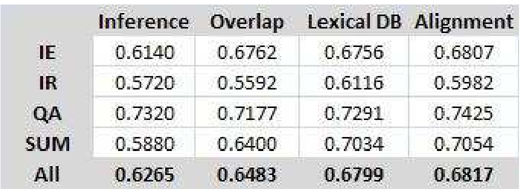
#	First Author (Group)	Run	Lexical DB	Overlap	Alignment	Inference
1	Adams (Dallas)	run1	x			
2	Bos (Rome & Leeds)	run1	x			
		run2	x			x
3	Burchardt (Saarland)	run1	x		x	
		run2	x		x	
4	Clarke (Sussex)	run1		x		
		run2		x		
5	de Marneffe (Stanford)	run1	x		x	x
		run2	x		x	x
6	Delmonte (Venice)	run1*	x		x	x
7	Ferrández (Alicante)	run1	x		x	
		run2	x		x	
8	Herrera (UNED)	run1	x			
		run2	x		x	
9	Hickl (LCC)	run1	x	x	x	
10	Inkpen (Ottawa)	run1	x	x	x	
		run2	x	x	x	
11	Katrenko (Amsterdam)	run1			x	
		run2		x		
12	Kouylekov (ITC-irst & Trento)	run1	x		x	
		run2	x		x	
13	Kozareva (Alicante)	run1	x	x		
		run2	x	x		
14	Litkowski (CL Research)	run1		x		
		run2			x	
15	Marsi (Tilburg & Twente)	run1	x		x	
16	Newman (Dublin)	run1	x	x		
		run2	x	x	x	
17	Nicholson (Melbourne)	run1	x		x	
		run2	x		x	
18	Nielsen (Colorado)	run1*		x	x	
		run2*		x	x	
19	Rus (Memphis)	run1	x		x	
		run2			x	
20	Schilder (Thomson & Minnesota)	run1	x		x	
		run2	x		x	
21	Tatu (LCC)	run1	x			x
22	Vanderwende (Microsoft Research & Stanford)	run1	x		x	
		run2	x		x	
23	Zanzotto (Milan & Rome)	run1	x		x	
		run2	x		x	

Table 5.6: The classification of the forty-one systems participating in RTE-2 into four major system types, based on their components description derived by (Bar-Haim et al., 2006).

cantly higher than **Inference** and **Overlap**.

On the other hand, the variance in performance is more moderate on the QA task, where all system types reach their highest performance. The best rate in this case is achieved by **Alignment** systems, though. It is remarkable that IR and QA are the only tasks where **Inference** achieves better results than **Overlap** (but not any other system type), which performs more poorly than any other system type on these two tasks.

All types but **Inference** seem to perform approximately in the same way on IE. Finally, **Lexical DB** and **Alignment** share the title of the best system types for SUM, with impressive performance rates of approx 0.7, as opposed to the 0.64 rate of **Overlap**, the second best system type of the task.



	Inference	Overlap	Lexical DB	Alignment
IE	0.6140	0.6762	0.6756	0.6807
IR	0.5720	0.5592	0.6116	0.5982
QA	0.7320	0.7177	0.7291	0.7425
SUM	0.5880	0.6400	0.7034	0.7054
All	0.6265	0.6483	0.6799	0.6817

Table 5.7: The distribution of average system performance of each of the four system types across the different tasks of RTE-2. The dataset covered is the positive entailment test set.

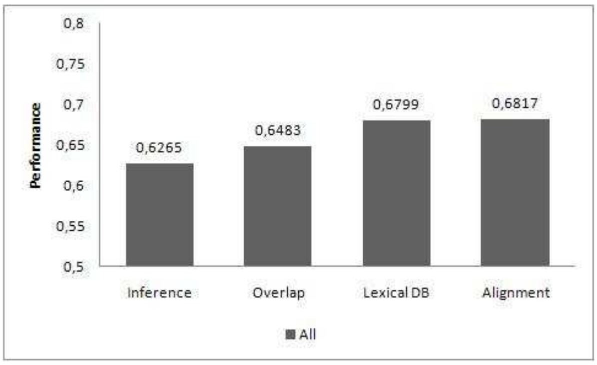


Figure 5.5: The chart of the distribution of average system performance of each system type in the overall positive entailment test set of RTE-2.

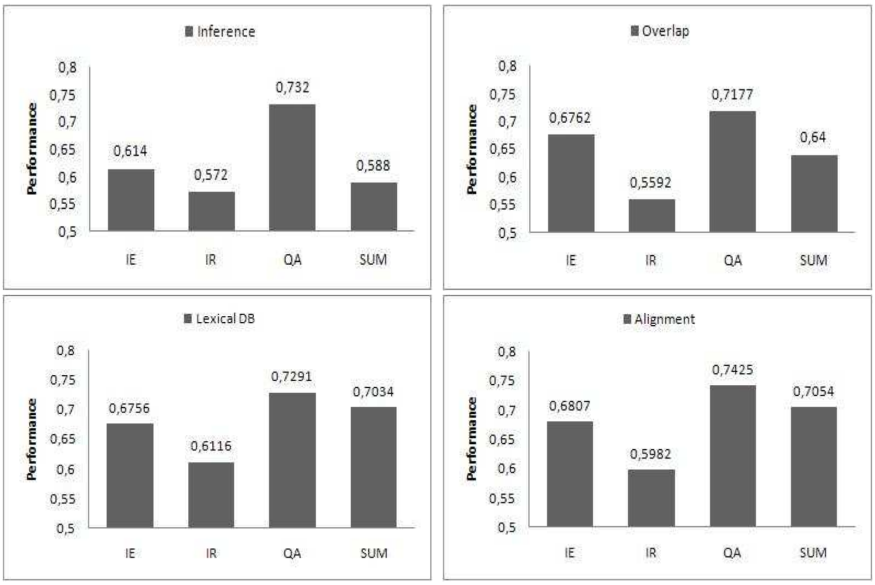
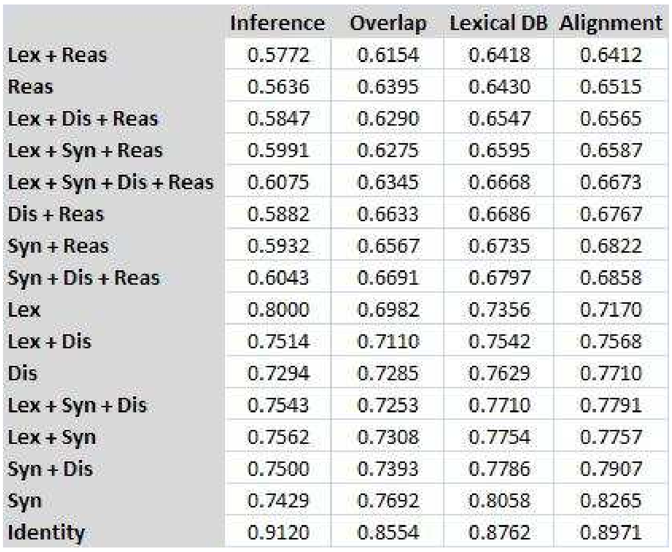


Figure 5.6: The charts of the distributions of average system performance of each system type individually across the tasks of RTE-2.

5.2.2 Performance Across Entailment Types

Even more interesting observations can be made by analyzing the performance of system types with respect to the different entailment types derived from the ARTE scheme annotation conducted. Thus, taking into account the entailment types of Table 4.8 of Section 4.2, we are in position to inspect the performance distributions of our four system types across the various entailment types in the datasets.

Table 5.8 summarizes these distributions, whereas Figures 5.7, 5.8, 5.9 and 5.10 display them separately for each system type. Entailment types here are meant in the *exclusive* sense, i.e. corresponding to the pairs that can be determined only with use of the entailment types indicated.



	Inference	Overlap	Lexical DB	Alignment
Lex + Reas	0.5772	0.6154	0.6418	0.6412
Reas	0.5636	0.6395	0.6430	0.6515
Lex + Dis + Reas	0.5847	0.6290	0.6547	0.6565
Lex + Syn + Reas	0.5991	0.6275	0.6595	0.6587
Lex + Syn + Dis + Reas	0.6075	0.6345	0.6668	0.6673
Dis + Reas	0.5882	0.6633	0.6686	0.6767
Syn + Reas	0.5932	0.6567	0.6735	0.6822
Syn + Dis + Reas	0.6043	0.6691	0.6797	0.6858
Lex	0.8000	0.6982	0.7356	0.7170
Lex + Dis	0.7514	0.7110	0.7542	0.7568
Dis	0.7294	0.7285	0.7629	0.7710
Lex + Syn + Dis	0.7543	0.7253	0.7710	0.7791
Lex + Syn	0.7562	0.7308	0.7754	0.7757
Syn + Dis	0.7500	0.7393	0.7786	0.7907
Syn	0.7429	0.7692	0.8058	0.8265
Identity	0.9120	0.8554	0.8762	0.8971

Table 5.8: The distributions of average performance of each system type across the different entailment types, in the *exclusive* sense.

A rather surprising fact is that the **Inference** system type, the par excellence type for success in deeper reasoning cases, performs worse than any other system type on all entailment types that involve **Reasoning**. This is particularly noteworthy, considering that the **Reasoning** entailment type is defined in a way that incorporates not only strict logical inferences, but also a large amount of background knowledge and paraphrase entailments (see Table 4.8 of Section 4.2), both of which are targets of the components employed by **Inference** systems.

On the contrary, as Figures 5.11 and 5.12 demonstrate, **Inference** outperforms all other system types for **Lexicon** and **Identity** entailments.

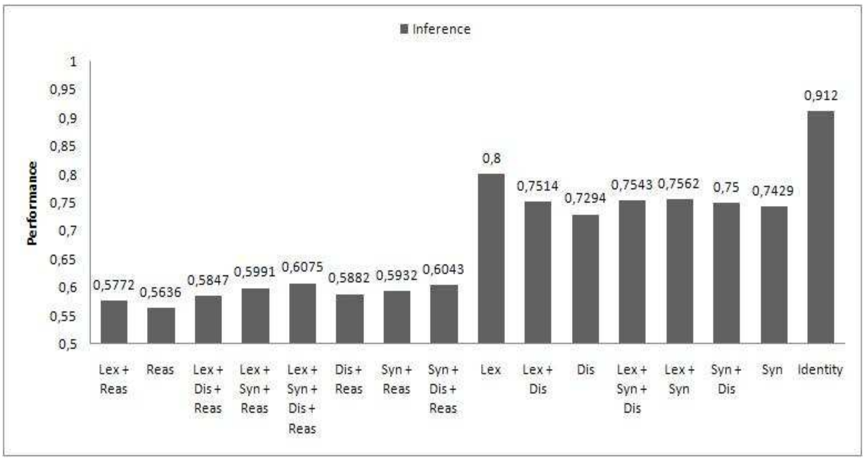


Figure 5.7: The chart of the average performance distribution of the **Inference** system type across the different entailment types, drawn by Table 5.8.

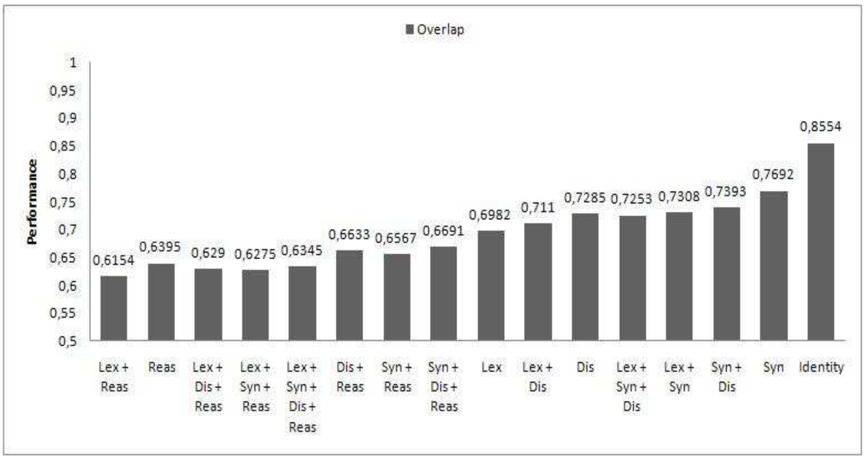


Figure 5.8: The chart of the average performance distribution of the **Overlap** system type across the different entailment types, drawn by Table 5.8.

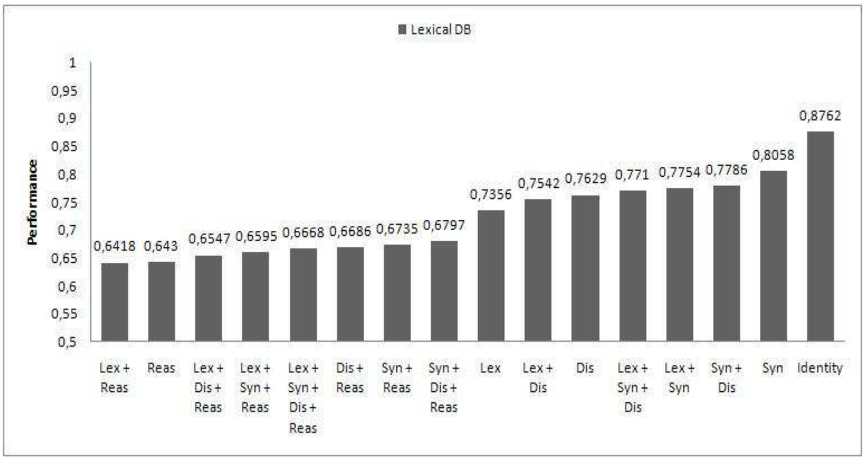


Figure 5.9: The chart of the average performance distribution of the **Lexical DB** system type across the different entailment types, drawn by Table 5.8.

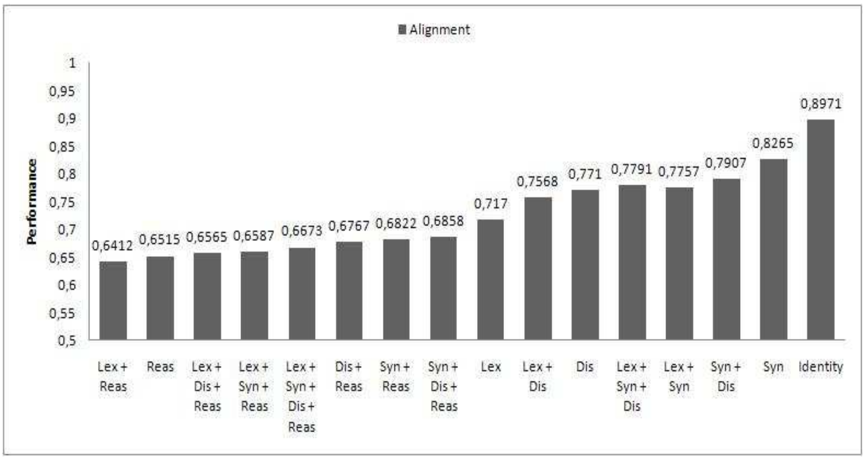


Figure 5.10: The chart of the average performance distribution of the **Alignment** system type across the different entailment types, drawn by Table 5.8.

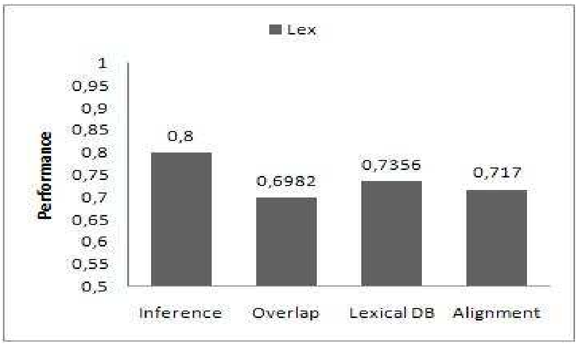


Figure 5.11: The chart of the distribution of average performance of the four system types on the **Lex** entailment type of Table 5.8.

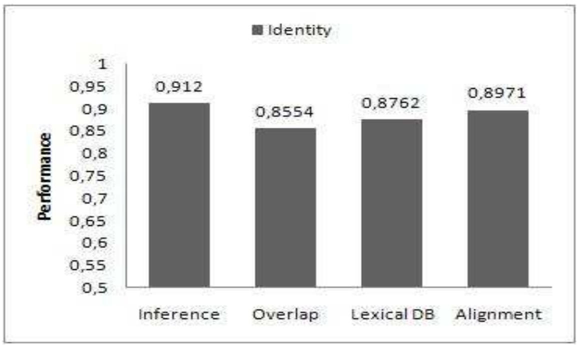


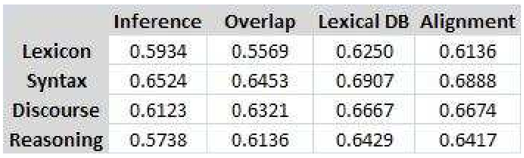
Figure 5.12: The chart of the distribution of average performance of the four system types on the **Identity** entailment type of Table 5.8.

These two entailment types are in fact the only ones for which the **Inference** system type achieves the best performance.

This leads us to the observation that there is no single entailment type for which **Overlap** systems achieve better performance than all other system types; on top of that, **Overlap** seems to be worse than all other system types for **Identity** entailments. Finally, **Lexical DB** and **Alignment** systems appear to be performing comparably well for practically all entailment types. **Lexical DB** slightly outperforms **Alignment** for **Lexicon** entailments, whereas the opposite is true for **Syntax-** and **Discourse**-related entailments.

Concluding this analysis, let us cast our eyes over the inclusive-sense classification of pairs into entailment types: Table 5.9 presents the distribution of average performance of system types across entailment types *inclusively.* That is, each entailment type here is associated with the pairs whose entailment value can be determined with use of the entailment type in question, possibly also combined with other entailment types. Figures 5.13, 5.14, 5.15 and 5.16 show the corresponding charts for each system type.

The results suggest that with the only exception of **Inference**, which demonstrates a minor variation, all system types exhibit the same pattern on performance across entailment types: **Lexicon** achieves the lowest performance rates, even for **Lexical DB** systems, followed by **Reasoning**. **Discourse** always receives higher rates of system performance, while, finally, **Syntax** appears as the "easiest" type of entailment for all the different types of systems.



	Inference	Overlap	Lexical DB	Alignment
Lexicon	0.5934	0.5569	0.6250	0.6136
Syntax	0.6524	0.6453	0.6907	0.6888
Discourse	0.6123	0.6321	0.6667	0.6674
Reasoning	0.5738	0.6136	0.6429	0.6417

Table 5.9: The distribution of average performance of each of the system types on the different entailment types of the positive entailment test subcorpus of RTE-2. Contrary to Table 5.8, the entailment types here are meant in the *inclusive* sense.

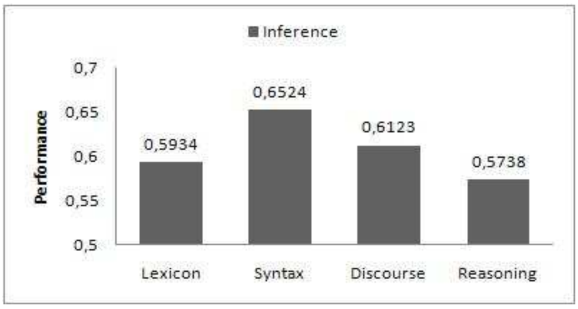


Figure 5.13: The chart of the distribution of average performance of the **Inference** system type across the different types of entailment, as in Table 5.9.

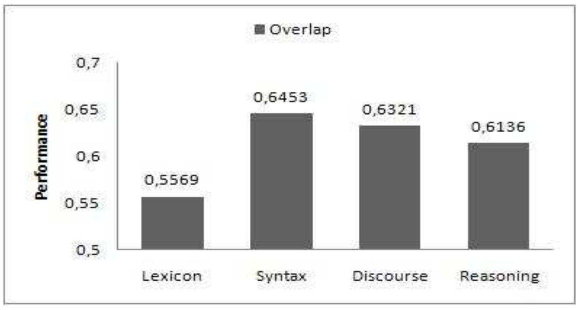


Figure 5.14: The chart of the distribution of average performance of the **Overlap** system type across the different types of entailment, as in Table 5.9.

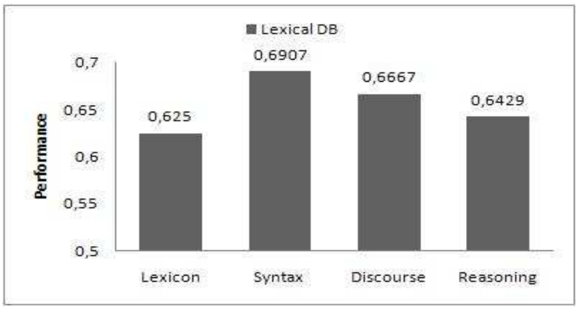


Figure 5.15: The chart of the distribution of average performance of the **Lexical DB** system type across the different types of entailment, as in Table 5.9.

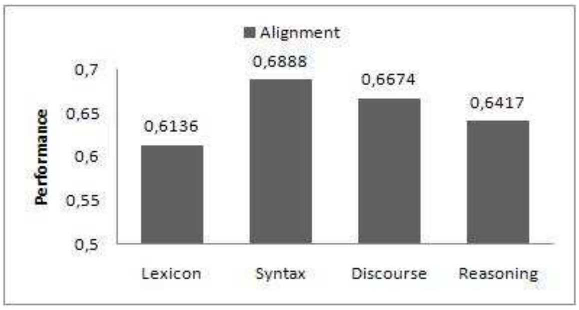


Figure 5.16: The chart of the distribution of average performance of the **Alignment** system type across the different types of entailment, as in Table 5.9.

5.2.3 Discussion

The classification of the RTE-2 systems into four types as proposed in Table 5.5 is a straightforward and intuitive way of distinguishing among differently built architectures. Grounded on the detailed overview of Bar-Haim et al. (2006), which covers a wide range of textual entailment mechanisms employed by the participating systems, it aims at providing a compact representation of the most characteristic among the system components utilized.

However, the choice of classifying the systems in this particular way is by no means the only possible, and perhaps even not the optimal for allowing for a categorization as discriminative as possible. Obviously, the classification adopted is rather crude and suffers from a large degree of overlap among the different system types.

The types **Lexical DB** and **Alignment**, for instance, are hard to differentiate, since an inspection of the resulting Table 5.6 reveals that the majority of the systems that belong to the one class are also to be found in the other. The instances of systems that belong exclusively to one of the four types are particularly limited: 3 for **Lexical DB**; 4 for **Overlap**; 3 for **Alignment**; 0 for **Inference**. On top of that, **Lexical DB** comprises all but 9 systems out of the 41. Finally, the classification does not account for the number of different components employed by the systems, which is arguably an interesting factor with regard to systems' performance.

Whereas an attempt at a more sophisticated and thoroughgoing classification of systems is beyond the scope of this study, we can still make certain interesting observations, if we survey the available classification from new perspectives.

Exclusiveness. One such perspective would take into account whether a given system type is the only one a system corresponds to, or one of a combination of different types, which all represent it.

Merging the information of Tables 5.1 and 5.6 from this perspective, we observe that the average overall accuracy of the 4 systems that function exclusively by overlap-based methods (system type **Overlap**) is fairly lower than the average accuracy of the 9 systems that use lexical overlapping in combination with other techniques (approx 0.56 vs. approx 0.59).

Oddly, the opposite is true for the **Lexical DB** type. The 3 systems that belong exclusively to the **Lexical DB** type achieve an average accuracy as high as approx 0.61, while the 29 systems that belong to this, in addition to at least one other system type, is approx 0.59.

On the other hand, we do not observe a significant difference in system accuracy for the **Alignment** type: The average accuracy of the 3 systems that use only **Alignment** is 0.58, whereas the incorporation of other components, realized by 27 systems, does not help it rise above approx 0.59.

The criterion of exclusiveness is not applicable to the **Inference** class, where none of the 5 system-members are found to rely on it exclusively. The average accuracy of the 5 **Inference** systems is approx at the 0.62 level, and thus higher than all other average rates of accuracy listed above. Given that **Inference** is the type with the worst performance on the positive entailment test set (see Table 5.7), this fact implies that **Inference** systems may handle negative entailments more successfully than positive ones.

Number of Components. Finally, another way of viewing Table 5.6 is by looking into how many different types correspond to each system.

Clearly, there is no system that is associated with all four different types; nevertheless, whether a system integrates one, two or three types appears to be producing a small difference in accuracy. Indeed, the average accuracy of the 10 systems that belong exclusively to one type—be it **Lexical DB**, **Overlap** or **Alignment**—is approx 0.58, equal to the one of the 24 systems that belong to exactly two different types. However, the 7 systems that belong to as many as three of the four different types achieve an average accuracy of approx 0.60, justifying the general confidence that combinations of variant techniques may produce better results.

5.3 Summary

In this chapter we advanced our understanding of the RTE-2 dataset by analyzing how real systems perform on it, both from a collective, and a system-type-specific perspective. The variant application settings and types of entailment were the parameters in the analysis.

Certainly the purpose of the study was not to exhaustively cover the topic, which is after all deeply related to the nature of the textual entailment phenomenon. The purpose was rather to set forth a methodology for seeing through the black box of textual entailment datasets and understanding the laws that govern it. In light of this, the present work can be considered as a novel step in this direction.

The main conclusions drawn are that system performance ranges significantly across the different datasets. Different system types also seem to behave differently for each input, although several limitations regarding the classification of systems into types have prevented us from drawing clear conclusions in this respect. The problem of learning about the relevant success of the different modules of textual entailment systems, and their optimized combinations, is not trivial, and calls for an extensive, fully fledged study.

Last but not least, a noteworthy discovery about the importance of the notion of overlap–entailment correlation in the dataset was made. Comparing it with the systems' performance notion for the different RTE-2 subcorpora, we observed that the behavior of the two is under all circumstances highly similar. This fact suggests that the correlation between overlap and entailment may be playing a crucial role in the RTE-2 datasets, whose appropriateness, as discussed in Subsection 5.1.2, is open to question.

Chapter 6

Conclusion

This chapter summarizes the main results of the thesis and outlines certain issues for further research raised by them.

6.1 Summary

The present work pursued the aim of a better understanding of applied textual entailment by means of examining textual entailment datasets and following clear methodological lines for their evaluation.

A novel scheme for annotation of such datasets, ARTE , was proposed. In this scheme the entailment problem is viewed in relation to three well-defined levels—Alignment, Context and Coreference—, each of which handles different aspects of the entailment mechanism. The scheme was found suitable for the data of RTE and was applied with success to a considerably large subcorpus of the RTE-2 test set. The resulting annotation enabled fresh insights into the semantic-linguistic properties of the textual entailment data and prepared the ground for an extensive analysis of their observed characteristics.

Shallow and deep aspects of the dataset were objects of this investigation. From the shallow perspective, covering word lengths, lexical overlap and the herein proposed notion of overlap–entailment correlation, we learnt about objective, external attributes of the dataset that can be directly measured and exploited by textual entailment systems.

From the deeper perspective, which realized a multifaceted statistical analysis of the distribution of the annotated entailment features, we made several discoveries with respect to the more intuitive, linguistic aspects of the data. It was noticed, for instance, that certain types of inference that play a major role in theoretical semantics, such as negation or counterfactive expressions, occur in the data very infrequently, or not at all. The proper detection, representation and quantification of such linguistic aspects is, nonetheless, a problem far from conclusively solved.

As a further step, the annotation results were put into service for a study of the relative performance of different existent textual entailment engines on datasets of variant nature. The submission results of the twenty-three research groups participating in the RTE-2 Challenge were analyzed both collectively and by being classified into broad classes, according to system component descriptions. In this analysis we came, for example, to the rather unexpected finding that systems using logical inference techniques apparently performed worse than other systems for reasoning-related entailments, but better than other systems for "simple" entailments based on surface-level identity.

The general conclusions drawn are that the subcorpora of the RTE-2 dataset seem to have significantly different characteristics, both at a surface, and at a deeper semantic level. The four individual application settings IE, IR, QA and SUM of RTE-2 appear to correspond to textual entailment tasks of genuinely different demands, which call for "customized" treatment. The varying rates of average system performance noticed across the different subcorpora provide further evidence for this. From the systems' performance study we make some first remarks about the efficiency of several techniques for recognition of textual entailment on a range of different data.

Finally, the study unveiled an interesting link between the notions of overlap–entailment correlation and systems' performance on the data. Impressively, the notion of overlap–entailment correlation seems capable of accounting for the performance of real systems much more satisfactorily than any other aspects of the analysis. This is a fact that puzzles over the actual suitability of the examined dataset as a means for the advancement of textual entailment technology.

6.2 Outlook

The work presented in this thesis introduced original ways for the evaluation of textual entailment datasets and demonstrated how they can be implemented for advancing our understanding of the mechanisms behind textual entailment, and the strategies for handling it. Though interesting results were produced in this way, the study does not claim to be complete. Further research needs to be placed in several directions before the full potential of the evaluation methodology proposed is uncovered.

The annotation scheme. The ARTE scheme is competent but certainly not yet fully perfected. Through accumulation of experience in textual entailment annotation, the scheme and the guidelines can improve and offer a more mature framework for evaluation. Especially the feature Reasoning can receive a more fine-grained analysis; also the annotation of non-entailment, which was merely an experimental undertaking, can be further attested on larger corpora, reviewed and expanded.

The annotation data. Textual entailment research can make good use of the annotated T–H corpora in ways extending beyond evaluation. For instance, useful linguistic patterns and entailment rules could be extracted and exploited by textual entailment systems. Moreover, the alignment information provided by the annotation can serve as a model for testing automatic system alignments.

The evaluation methodology. In a larger-scale study the analysis of the annotated data could and should be conducted in a more systematic way. Tools only slightly explored here, such as data clustering, can be applied more exhaustively. Additionally, the classification of entailment features into types of entailment, and system components into system types, requires a more thorough investigation in order to enable reliable and useful conclusions about the data.

Case studies. A particularly interesting potential of the annotation lies in the direction of employing it for direct error analysis and diagnosis of strengths and weaknesses of individual textual entailment systems. A system could clearly gain benefit from learning which entailment types it handles with success, as well as on which entailment types it produces wrong positive or wrong negative answers.

The compilation of textual entailment datasets. The annotated sample finally can help the research community set the standards for better-controlled textual entailment datasets, which can advance the field by fixing more carefully determined goals for the state-of-the-art textual entailment engines.

Bibliography

Baker, C., Fillmore, C., and Lowe, J. (1998). *The Berkeley Framenet project.* In Proceedings of the COLING-ACL, Montreal, Canada.

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., and Szpektor, I. (2006). *The Second PASCAL Recognising Textual Entailment Challenge.* In Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, Venice, Italy.

Bar-Haim, R., Szpektor, I., and Glickman, O. (2005). *Definition and Analysis of Intermediate Entailment Levels.* In Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, 55–60. Ann Arbor.

Bayer, S., Burger, J., Ferro, L., Henderson, J., and Yeh, A. (2005). *MITRE's Submissions to the EU Pascal RTE Challenge.* In Proceedings of the PASCAL Challenges Workshop on Recognizing Textual Entailment, 41–44. Southampton, U.K.

Chierchia, G. and McConnell-Ginet, S. (2000). *Meaning and grammar (2nd ed.): An introduction to semantics.* MIT Press, Cambridge, MA, USA.

Clark, P., Murray, W. R., Thompson, J., Harrison, P., Hobbs, J., and Fellbaum, C. (2007). *On the Role of Lexical and World Knowledge in RTE3.* In Proceedings of the Workshop on Textual Entailment and Paraphrasing, 171–176. Prague.

Crouch, R., Karttunen, L., and Zaenen, A. (2006). *Circumscribing is not excluding: A response to Manning.* Unpublished manuscript. http://www2.parc.com/istl/members/karttune/publications/reply-to-manning.pdf.

Crouch, R., Sauri, R., and Fowler, A. (2005). *AQUAINT Pilot Knowledge-Based Evaluation: Annotation Guidelines.* Unpublished manuscript. http://www2.parc.com/istl/groups/nltt/papers/aquaint_kb_pilot_evaluation_guide.pdf.

Dagan, I. and Glickman, O. (2004). *Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability.* In Proceedings of the PASCAL Workshop on Learning Methods for Text Understanding and Mining. Grenoble, France.

Dagan, I., Glickman, O., and Magnini, B. (2006). *The PASCAL Recognising Textual Entailment Challenge.* In Quiñonero-Candela et al. (Eds.): MLCW 2005, LNAI Volume 3944, 177–190. Springer-Verlag.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database.* The MIT Press, Cambridge, MA.

Fillmore, C. J., Baker, C. F., and Sato, H. (2002). *Seeing Arguments through Transparent Structures.* In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC). Las Palmas. 787–791.

Giampiccolo, D., Magnini, B., Dagan, I., and Dolan, B. (2007). *The Third PASCAL Recognising Textual Entailment Challenge.* In Proceedings of the Workshop on Textual Entailment and Paraphrasing, 1–9. Prague.

Glickman, O. (2006). *Applied Textual Entailment.* Ph.D. Thesis. Bar Ilan University.

Glickman, O., Dagan, I., and Koppel, M. (2005). *Web Based Probabilistic Textual Entailment.* In Proceedings of the PASCAL Challenges Workshop on Recognizing Textual Entailment, 41–44. Southampton, U.K.

Grice, P. (1975). *Logic and conversation.* In P. Cole, ed., Syntax and Semantics. Volume 3, 41–58. New York: Academic Press.

Karttunen, L. and Zaenen, A. (2005). *Veridicity.* In Annotating, Extracting and Reasoning about Time and Events. Dagstuhl Seminar Proceedings 05151. Dagstuhl, Germany.

Kroeger, P. (2005). *Analyzing Grammar: An Introduction.* Cambridge University Press.

Levinson, S. (1983). *Pragmatics.* Cambridge, England: Cambridge University.

MacCartney, B., Grenager, T., de Marneffe, M.-C., Cer, D., and Manning, C. D. (2006). *Learning to recognize features of valid textual entailments.* In Proceedings of HLT-NAACL-06.

Manning, C. D. (2006). *Local textual inference: It's hard to circumscribe, but you know it when you see it—and NLP needs it.* Unpublished manuscript. http://nlp.stanford.edu/˜manning/papers/LocalTextualInference.pdf.

Manning, C. D., Moldovan, D., and Voorhees, E. (2007). *Annotation guidelines for marking contradictions.* Unpublished manuscript. http://nlp.stanford.edu/RTE3-pilot/contradictions.pdf.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). *Introduction to WordNet: An On-line Lexical Database.* In International Journal of Lexicography 3 (4). Revised August 1993.

Müller, C. and Strube, M. (2006). *Multi-Level Annotation of Linguistic Data with MMAX2.* In Braun, S., Kohn, K., and Mukherjee, J. (Eds.): Corpus Technology and Language Pedagogy. New Ressources, New Tools, New Methods. Frankfurt: Peter Lang, 197–214. (English Corpus Linguistics, Vol.3).

Nairn, R., Condoravdi, C., and Karttunen, L. (2006). *Computing relative polarity for textual inference.* In Proceedings of ICoS-5. Buxton, UK.

Och, F. J. and Ney, H. (2003). *A systematic comparison of various statistical alignment models.* Computational Linguistics, 29(1).

Sag, I. A. (1997). *English Relative Clause Constructions.* In Journal of Linguistics. Volume 33, 431–484.

Schmid, H. (1994). *Probabilistic part-of-speech tagging using decision trees.* In Proceedings of International Conference on New Methods in Language Processing, Manchester, U.K.

Vanderwende, L., Coughlin, D., and Dolan, B. (2005). *What Syntay can Contribute in Entailment Task.* In Proceedings of the PASCAL Challenges Workshop on Recognizing Textual Entailment, 13–16. Southampton, U.K.

Wang, R. and Neumann, G. (2007). *Recognizing Textual Entailment Using a Subsequence Kernel Method.* In Proceedings of AAAI 2007, 937–942. Vancouver, Canada.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques (2nd ed.).* Morgan Kaufmann, San Francisco.

Zaenen, A., Karttunen, L., and Crouch, R. (2005). *Local Textual Inference: Can it be defined or circumscribed?* In Proceedings of the ACL 2005 Workshop on Empirical Modelling of Semantic Equivalence and Entailment. 31–36.

Appendix A

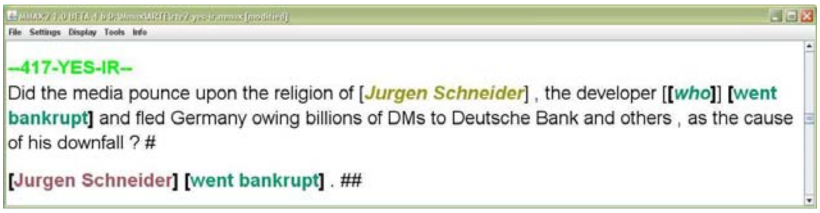
The ARTE Guidelines

The following sections complement Chapter 3 in the description of the ARTE scheme, our annotation scheme for RTE. They provide definitional details of the concepts involved and concrete guidelines on their application. In parallel they address some of the technical issues arising.

A.1 Annotation Basics

The annotation platform is the MMAX2 tool (Müller and Strube, 2006). The data consist of T–H pairs, where H is either entailed by T or not, according to a gold standard. H is a (usually short) single sentence, and T consists of one or two sentences, as described in (Bar-Haim et al., 2006). Each pair is preceded by an ID code and is followed by a horizontal line, separating it from the next pair. A single hash # marks the end of each T, while two hashes ## stand at the point where each H ends.

It is noteworthy that T may involve questions and not statements, since these can also carry presuppositions conveying information, as in Figure A.1.



--417-YES-IR--

Did the media pounce upon the religion of [*Jurgen Schneider*] , the developer [[***who***]] [**went bankrupt**] and fled Germany owing billions of DMs to Deutsche Bank and others , as the cause of his downfall ? #

[**Jurgen Schneider**] [**went bankrupt**] . ##

Figure A.1: A rather not typical pair, where T is in the form of a question. It contains however a relative clause triggering a presupposition, on which the entailment is based.

The relevant parts of the text are marked and assigned specific values of a predefined list of corresponding attributes and relations. This way selected fragments in the T–H pair are placed in a particular relationship, and/or receive specific labels, which provide a characterization of their role with respect to the entailment. These selected fragments are called **markables**, and may be *discontinuous*, as well as *overlapping*. Each markable is by default inextricably accompanied by a certain list of **features** describing it. The features supported are of two types:

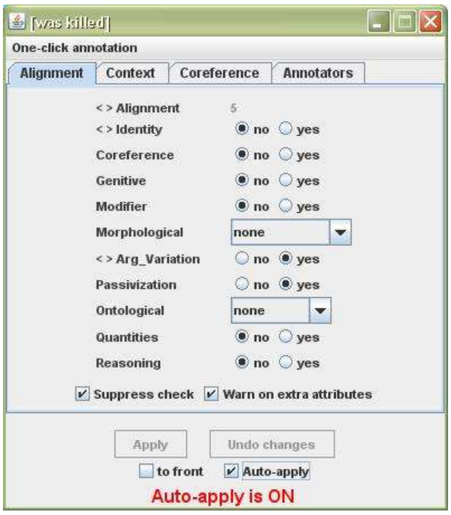
Attributes. Attributes consist of a name and a set of possible values, one of which is always selected. Possible values are displayed as either a number of radio buttons or as a drop-down list. In some cases the availability of an attribute depends on the current value of another; then the former is called **dependent** and the latter **branching** attribute. Finally a special type of attribute is the free text field, which can accept any string as its value.

Relations. Apart from carrying annotations in the form of attribute-value pairs, a markable can also be associated with (one or more) other markables to form markable relations. The type of relations we use are **pointer** relations, which always associate with one markable (the **source**) one or more **target** markables in an intransitive, directed fashion.

The annotation resides in XML file format, but is also visualized with the help of two separate windows. The data appear in the **main window**, as presented in Figures 3.3 and A.1, where the markables are rendered in varying styles and colors. Markables are sensitive to mouse clicks; once selected, a markable is highlighted and its corresponding list of features is displayed in the **attribute window**, as presented in Figure A.2. If the selected markable is participating in a relation as source, this relation will also be visualized in the main window by means of an arc linking source and target.

The annotation is structured in several distinct levels (layers), with different functions and purposes each: Alignment, Context, Coreference and Annotators[1] level for positive entailment; Non-entailment and Annotators level for negative entailment.

[1] The Annotators level was not mentioned in Chapter 3 because its use serves exclusively the annotation procedure and is not associated with the annotation itself.



< > Alignment 5
< > Identity ● no ○ yes
Coreference ● no ○ yes
Genitive ● no ○ yes
Modifier ● no ○ yes
Morphological none
< > Arg_Variation ○ no ● yes
Passivization ○ no ● yes
Ontological none
Quantities ● no ○ yes
Reasoning ● no ○ yes
☑ Suppress check ☑ Warn on extra attributes

□ to front ☑ Auto-apply
Auto-apply is ON

Figure A.2: The attribute window, containing the list of attributes associated with an Alignment-markable.

A.2 Alignment

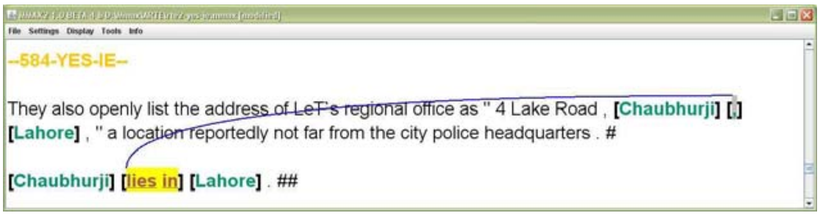
As described in Subsection 3.1.1, the Alignment level is the most basic of the positive entailment annotation. Every fragment of H should be aligned to at least one corresponding part of T. Whenever such an alignment is made, it is automatically assigned a label specifying its nature, meant to be appropriately modified in the annotation process.

A.2.1 Markables

H is scanned and each and every single word in it becomes part of at least one markable, not excluding punctuation marks. The basic idea behind the selection of alignment markables in H is that they generally correspond to syntactic constituents of the sentences, although this claim is more of an intuitive rather than literal nature. Thus, in their majority they can be classified under the following basic syntactic categories: V-, NP-, AdjP-, Adv- and PP-markables.

However, it is not uncommon that we have other types of markables, such as Adj- or P-markables, depending on the specifics of the inference. A special case arises when the inference is heavily based on a non-trivial use of punctuation; then a punctuation-markable is allowed, as in Figure A.3.

The way markables are selected at this level differs between T and H in several aspects. One crucial difference is that, contrary to H, the words of



--584-YES-IE--

They also openly list the address of LeT's regional office as " 4 Lake Road , [Chaubhurji] [,] [Lahore] , " a location reportedly not far from the city police headquarters . #

[Chaubhurji] [lies in] [Lahore] . ##

Figure A.3: The information conveyed in the predicate of H is based on the appropriate interpretation of a comma in T, and therefore this comma will constitute the corresponding Alignment-markable.

T do not need to exhaustively be parts of markables. The markables of T are selected in a more eclectic way and on the grounds of their contribution to the entailment.

Another difference involves the way NP-markables are formed. The notion of **determiner** for creating the NP-markables in T is restricted to the traditional classes of articles (i.e., *a*, *an*, *the*), quantifiers[2] and demonstrative pronouns (e.g., *this*, *that*). Since coreference resolution in T is treated separately at the Coreference level, possessives in T—either in the form of pronouns (e.g., *his*, *whose*), or in the form of noun phrases (e.g., *Tibet's*)—will not constitute NP-markables together with the nouns they modify, but rather stand-alone markables, if required. On the other hand, coreference resolution in H is not addressed. Therefore possessives in H will also normally constitute part of the NP-markables.

Nonetheless in both H and T, what ultimately determines how large the markables should be and where the boundaries among them are to be drawn is not constituency but rather the nature of the inference mechanisms employed.

General heuristics

More specifically, markable creation for the Alignment level is a recursive process guided by both semantic and syntactic principles. It typically follows this pattern:

1. We start by examining the main clause of H.

2. We identify the main predicate. If it is the copula, then it is made one markable together with the predicate complement; otherwise it

[2] The notion of **quantifier** encompasses indefinite pronouns (e.g., *all*, *many*, *some*) and cardinal—but not ordinal—numbers.

constitutes one separate markable. We include any auxiliary verbs in this markable.

3. The subject of the main predicate is marked as a separate markable. Coordinate NPs (i.e., NPs linked by conjunctions such as *and* and *or*) do not need to be split.

4. The subordinate clauses of the main predicate are identified. If they are NP complements or adverb/PP adjuncts, they are made separate markables.

5. Punctuation marks are also included in markables, which may normally be selected arbitrarily from among the neighboring markables. If, however, the punctuation plays a particularly significant role in establishing the entailment, it may constitute a markable of its own. The end full-stop of the sentence may, as a single exception, remain unmarked and not belong to any markable.

6. For each remaining subordinate clause (e.g., clausal complements, adverbial clauses, relative clauses) we repeat steps 2–5, putting the subordinating conjunction in one markable together with the main predicate of the subordinate clause.

7. Once exhaustively arranged the words of H into markables, we turn to T.

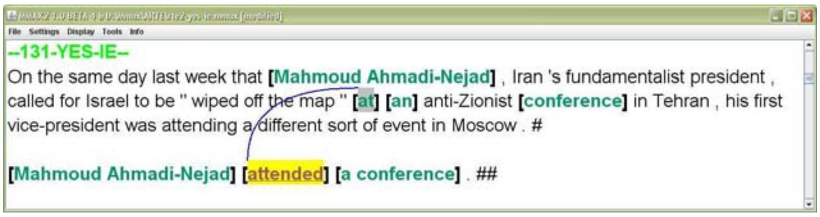
8. In T we look for the corresponding parts of the existent markables, i.e, for the pieces that justify the truth of H. We start by identifying the fragments of T that correspond to V-markables of H; then the other markables of H are examined. The fragments that can be matched to the H-markables should constitute the T-markables. Here the basic consideration is semantic and not syntactic, although the syntactic constituency is respected as far as possible.

Divergences

It must be pointed out that the account of markable creation given in the previous paragraph only covers the cases in which each of the constituents of H, as they were described, can directly be aligned to corresponding constituents in T. Clearly, in order to present as fine-grained and informative alignments as possible, this cannot always be the case.

As an example, it may be required that a single NP constituent consisting of Det + Adj + N be split into two separate subconstituents Det + N and

Adj, because the matching with parts of T can be modeled more accurately this way. Furthermore, a preposition, which would normally belong together with its PP, could be considered separately, as in Figure A.4, or even as part of the predicate (e.g., in case of a phrasal verb), if its semantic interpretation calls for it. Finally Figure A.5 provides an example of a predicate and its subject forming together a single markable.

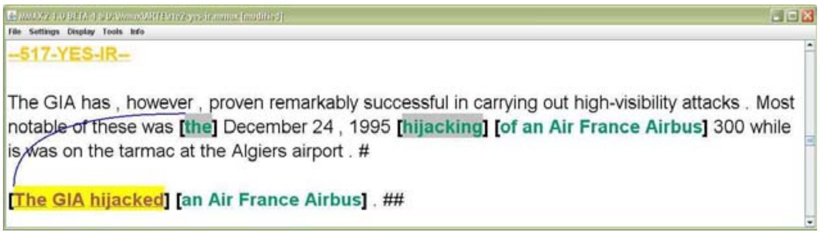


--131-YES-IE--

On the same day last week that [Mahmoud Ahmadi-Nejad] , Iran 's fundamentalist president , called for Israel to be " wiped off the map " [at] [an] anti-Zionist [conference] in Tehran , his first vice-president was attending a different sort of event in Moscow . #

[Mahmoud Ahmadi-Nejad] [attended] [a conference] . ##

Figure A.4: The V-markable *attended* aligns to the preposition *at* and not to the PP it is part of. Additionally, the NP *an anti-Zionist conference* is stripped off its adjective, leaving only the remaining phrase *an . . . conference* as a discontinuous markable, since this is the only relevant part for the entailment.



--517-YES-IR--

The GIA has , however , proven remarkably successful in carrying out high-visibility attacks . Most notable of these was [the] December 24 , 1995 [hijacking] [of an Air France Airbus] 300 while is was on the tarmac at the Algiers airport . #

[The GIA hijacked] [an Air France Airbus] . ##

Figure A.5: The agent of the *hijacking* mentioned in T appears in a different sentence and can only be inferred by the context. For this reason the subject of the predicate *hijacked* in H will not be aligned separately, but will form a single markable with the predicate.

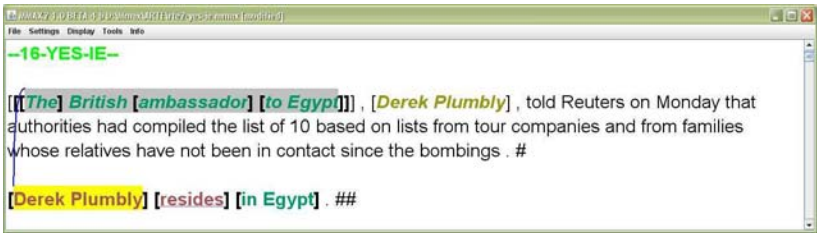
Conclusively, whenever syntactic constituency as described and semantic analysis are not in perfect agreement, the latter is the one that carries more weight and drives the process of markable creation.

Aligning appositives

One particular configuration frequently appearing in T involves appositive constructions, where two coreferential elements—normally NPs—are placed side by side, either separated by punctuation markers like commas or parentheses, or not. Under such circumstances it is not always straightforward to determine which of the two elements should constitute a markable at the Alignment level to match a coreferential markable of H.[3] Therefore the following principles are adopted:

Case 1: Punctuation marks the appositive. Then the component selected as a T-markable is the one unmarked by punctuation, and it is made a markable intact as an NP, including even modifiers or adjuncts that are not relevant for the matching and would normally have been stripped off it. Figure A.6 provides an example.

Case 2: The appositive is not marked by punctuation. In this case selected as T-markable is any component that directly matches the corresponding H-markable as normally, overlooking the appositive construction. Figure A.7 presents such a case.



--16-YES-IE--

[[[The] British [ambassador] [to Egypt]]] , [Derek Plumbly] , told Reuters on Monday that authorities had compiled the list of 10 based on lists from tour companies and from families whose relatives have not been in contact since the bombings . #

[Derek Plumbly] [resides] [in Egypt] . ##

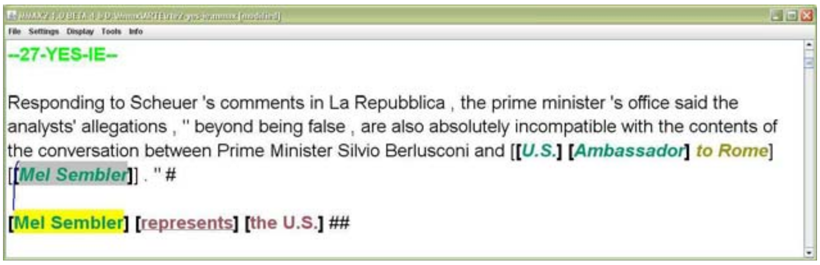
Figure A.6: Here the appositive NP *Derek Plumbly* is surrounded by commas. Therefore the NP selected as an Alignment-markable is the other one, even though *Derek Plumbly* appears identically in H.

A.2.2 Features

Figure A.2 presented the list of features used at the Alignment level. The markables of this level are exclusively designed to become members of ordered pairs, which participate in an alignment relation between T and H. This means that the basic feature here will be the pointer relation Alignment, directed from H-markables to T-markables. Once set, the Alignment relation will unfold a list of dependent features[4] to label it, as follows:

[3] In other words, which will be the "prominent" part of the appositive, as explained in Subsection A.4.2.

[4] To be precise, Alignment has only one dependent feature: Identity. But if Identity is *not* selected, then the complete set of the rest of the features is available.



--27-YES-IE--

Responding to Scheuer 's comments in La Repubblica , the prime minister 's office said the analysts' allegations , " beyond being false , are also absolutely incompatible with the contents of the conversation between Prime Minister Silvio Berlusconi and [[U.S.] [Ambassador] to Rome] [[Mel Sembler]] . " #

[Mel Sembler] [represents] [the U.S.] ##

Figure A.7: Here the NP *Mel Sembler* is not preceded by any punctuation mark. Therefore it is selected as an Alignment-markable of T to match its identical counterpart in H.

Identity. Only content words are considered (e.g., *the Kyodo news agency* ⟷ *Kyodo news agency*) and, normally, identity only at the base-form level of those is sufficient; e.g. *testifying* ⟷ *testified*. Particles and subordinating/coordinating conjunctions are also ignored, as Figure 3.4 indicated, unless they make a heavy contribution to the semantics of the markable.

Nonetheless lexical identity is not a sufficient condition for an Identity alignment, since the annotation takes the point of view of a human interpreting the sentences to judge the entailment, and not of a lexical matching system. For example, in Figure A.8 the same markable performs substantially different grammatical functions in T and H (of modifier and subject, respectively), and for this reason it is best labeled not as Identity, but as Modifier.

Note that this fact does not imply that two identically aligned markables must have the same grammatical function. However, if one of the two more informative features defined—Genitive and Modifier—apply, then the alignment is labeled as such and not merely as Identity.

Coreference. Following (Kroeger, 2005), we make a terminological distinction between relative pronouns (e.g., *who*, *which*) and the relativizer *that*, but the feature applies to both. It is used in case the coreference between T and H is established due to the linking of the T-markable to some other part of T, outside the markable boundaries, and therefore made evident because of the context in which the T-markable appears. This opens up two different possibilities.

1. Context coreferent. The text fragment of T outside the markable which establishes the coreference is an NP, and is coreferent to the

aligned markables; In this case the Coreference level is also active and captures this instance, as Figures 3.3 and 3.12 illustrated.

Note that in this case, if the Coreference level provides us with an anaphoric—as opposed to supplemental— relation linking the T-markable with its context, then the selection of the T-markable will involve the whole NP, including any modifiers or complements of the head noun. However relative clauses or appositives will be left out.

2. Context not coreferent. The context establishing the coreference of the markables is not actually a coreferent NP, but merely a related phrase in the text, which provides the necessary information. Figure A.9 illustrates this. In this case, contrary to the first, the annotation lives only at the Alignment level and does not extend to the Coreference level, as the only actual coreference has already been captured by the alignment.

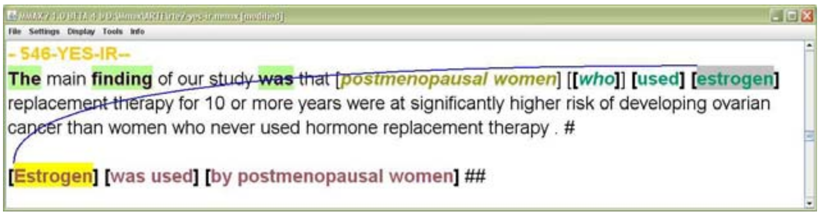
For simplicity reasons, Coreference alignments are made exclusively for NPs of H which are linked to another NP or any anaphoric expression in T, and not for anaphoric expressions in H, like pronouns. Such expressions in H either align to identical parts of T, if such an alignment is possible, or they merely form part of their predicate's markable and are not taken into account for the alignment of the markable, as Figure A.10 shows.

Genitive. Apart from the traditional genitive constructions, this feature also applies to the case of **transparent** noun (N) constructions, i.e., constructions of the form N *of* N, in which the first N is transparent[5] with regard to selectional relations between the second N and the external context. In these cases the transparent N will not constitute part of the markable, as indicated in Figure A.11.

However the Genitive feature is not uniformly used for all instances of genitive case. There are certain cases where its use is not considered purposeful:

1. An alignment is not marked as Genitive if it does not stand next to an alignment—either direct or indirect via a Coreference level

[5] In the sense of there being a discrepancy between the syntactic and the semantic head of the structure. As Fillmore et al. (2002) suggest, nouns that behave in this way are of several kinds; for example nouns denoting parts (e.g., *part of the room*), measures (e.g., *liter of wine*), units (e.g., *bout of the flu*), types (e.g., *kind of fish*), etc.

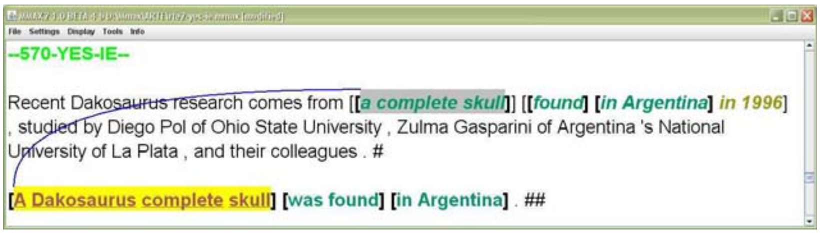


- 546-YES-IR--

The main finding of our study was that [postmenopausal women] [[who]] [used] [estrogen] replacement therapy for 10 or more years were at significantly higher risk of developing ovarian cancer than women who never used hormone replacement therapy . #

[Estrogen] [was used] [by postmenopausal women] ##

Figure A.8: Here the two lexically identical markables are not aligned with an Identity, but with a Modifier label.

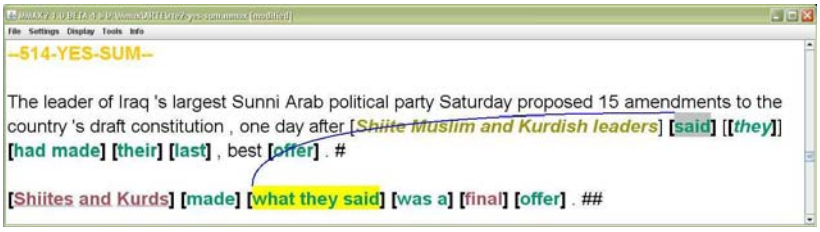


--570-YES-IE--

Recent Dakosaurus research comes from [[a complete skull]] [[found] [in Argentina] in 1996] , studied by Diego Pol of Ohio State University , Zulma Gasparini of Argentina 's National University of La Plata , and their colleagues . #

[A Dakosaurus complete skull] [was found] [in Argentina] . ##

Figure A.9: An alignment marked as Coreference + Reasoning. There is no coreferent context here.

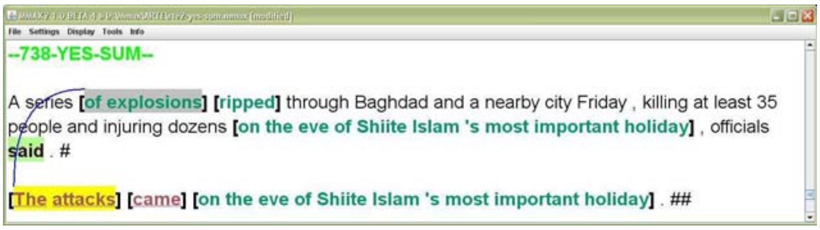


--514-YES-SUM--

The leader of Iraq 's largest Sunni Arab political party Saturday proposed 15 amendments to the country 's draft constitution , one day after [Shiite Muslim and Kurdish leaders] [said] [[they]] [had made] [their] [last] , best [offer] . #

[Shiites and Kurds] [made] [what they said] [was a] [final] [offer] . ##

Figure A.10: Even though the reference resolution of the personal pronoun *they* of H is important for the entailment, it simply takes part in an Identity alignment, since anaphoric expressions in H are not analyzed.



--738-YES-SUM--

A series [of explosions] [ripped] through Baghdad and a nearby city Friday , killing at least 35 people and injuring dozens [on the eve of Shiite Islam 's most important holiday] , officials said . #

[The attacks] [came] [on the eve of Shiite Islam 's most important holiday] . ##

Figure A.11: The noun *series* in T is transparent in the construction *series of explosions* and therefore is left out of the T-markable aligning to the H-markable *the attacks*. The alignment is labeled as Genitive + Reasoning.

link—between the head of the genitive-markable and a head or dependent of the other markable, as Figure A.12 exemplifies.

2. The Genitive feature is not selected for alignments on which both markables are in genitive case and their heads are aligned, as in Figure A.13.

3. In the trivial case of a participation in an action, where the genitive simply indicates the agent or patient (**subjective** or **objective** genitive, respectively), the alignment is not labeled as Genitive. Such an example is provided in Figure A.14. This applies also when the markables pair denoting the action is not a pair of a verb and its corresponding verbal noun, as in Figure A.14, but a verb–noun pair linked in a *straightforward* way, as in Figure A.15.

   We note that by **straightforward** alignment, we mean one that is not labeled as Reasoning. This notion should not be confused with the one of *multilabeled* alignment, which is discussed in Subsection A.2.3.

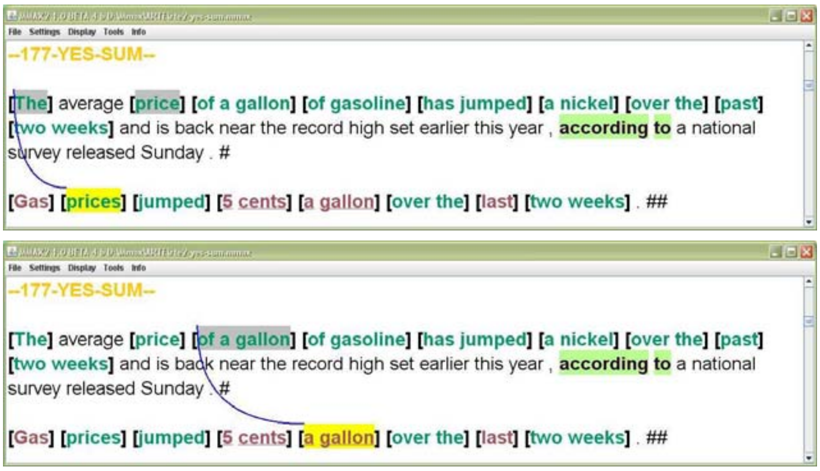
4. Finally, when the Genitive feature is selected, we do not additionally select Reasoning, since Genitive is designed especially to model the type of reasoning associated with genitive constructions. However when the reasoning involved is of another type, as in Figure A.16, the Reasoning feature is also selected. Moreover, labels indicating additional relations (e.g., Morphological or Ontological) may freely be selected.

Modifier. The use of the Modifier feature is guided by similar principles as the ones applicable to Genitive:

1. Similarly to Genitive, an alignment cannot be labeled as Modifier if there is no alignment between what is modified by the modifier-markable, and a head or dependent of the corresponding markable.

2. An alignment is not marked as Modifier in case both markables are in modifier position and their heads are in alignment, as Figure A.17 illustrates.

3. In parallel to Principle 3. for Genitive, Modifier does not apply to "trivial" cases, denoting merely the participants of actions, as in Figure A.18.



--177-YES-SUM--

[The] average [price] [of a gallon] [of gasoline] [has jumped] [a nickel] [over the] [past] [two weeks] and is back near the record high set earlier this year , according to a national survey released Sunday . #

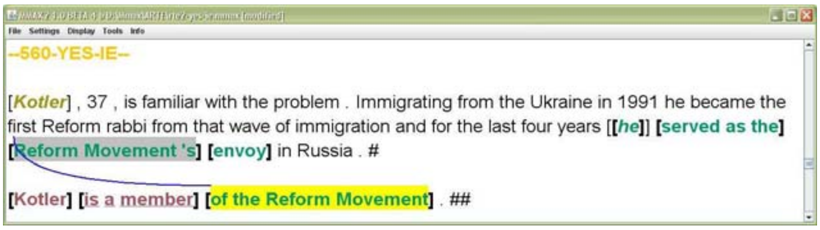
[Gas] [prices] [jumped] [5 cents] [a gallon] [over the] [last] [two weeks] . ##

--177-YES-SUM--

[The] average [price] [of a gallon] [of gasoline] [has jumped] [a nickel] [over the] [past] [two weeks] and is back near the record high set earlier this year , according to a national survey released Sunday . #

[Gas] [prices] [jumped] [5 cents] [a gallon] [over the] [last] [two weeks] . ##

Figure A.12: The T-markable *of a gallon* is in genitive case, but its corresponding H-markable *a gallon* is not in a direct dependency relation to the H-markable *prices*, which aligns to the genitive-markable's head *the price*. Therefore the Reasoning and not the Genitive feature will apply here.

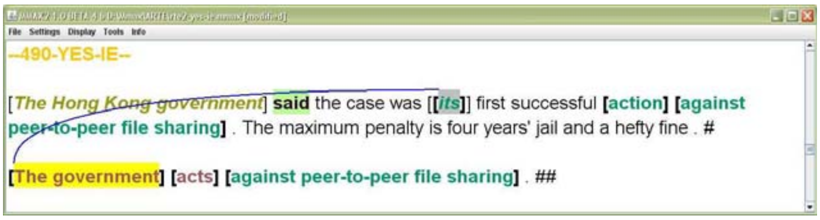


--560-YES-IE--

[*Kotler*] , 37 , is familiar with the problem . Immigrating from the Ukraine in 1991 he became the first Reform rabbi from that wave of immigration and for the last four years [[*he*]] [served as the] [Reform Movement 's] [envoy] in Russia . #

[Kotler] [is a member] [of the Reform Movement] . ##

Figure A.13: This alignment is marked as Identity and not as Genitive, since both markables are in genitive case and their heads are in alignment.

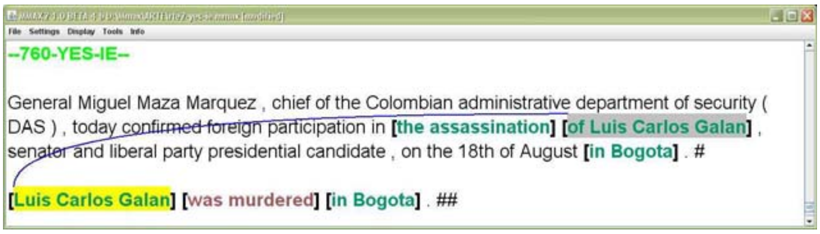


--490-YES-IE--

[*The Hong Kong government*] said the case was [[*its*]] first successful [action] [against peer-to-peer file sharing] . The maximum penalty is four years' jail and a hefty fine . #

[The government] [acts] [against peer-to-peer file sharing] . ##

Figure A.14: The alignment is between the subject of a verbal noun and a subjective possessive pronoun. Therefore it is marked exclusively as Coreference, and not as Genitive.

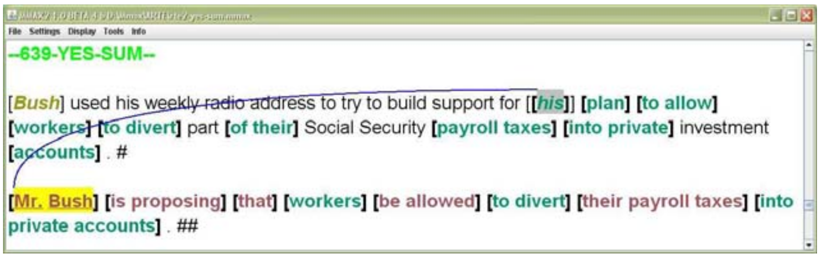


--760-YES-IE--

General Miguel Maza Marquez , chief of the Colombian administrative department of security (DAS) , today confirmed foreign participation in [the assassination] [of Luis Carlos Galan] , senator and liberal party presidential candidate , on the 18th of August [in Bogota] . #

[Luis Carlos Galan] [was murdered] [in Bogota] . ##

Figure A.15: Here the markables *was murdered* and *the assassination* express the same action and are linked with Synonymy + Nominalization, since the verbs *murder* and *assassinate* are synonymous. Therefore the genitive case expressed by the T-markable *of Luis Carlos Galan* is considered as objective, and the Genitive feature is not applied to its alignment in H.

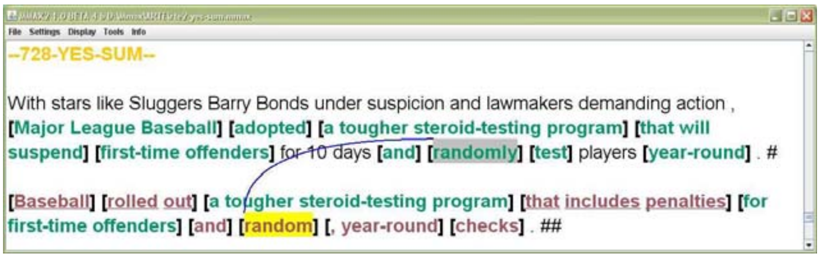


--639-YES-SUM--

[Bush] used his weekly radio address to try to build support for [[his]] [plan] [to allow] [workers] [to divert] part [of their] Social Security [payroll taxes] [into private] investment [accounts] . #

[Mr. Bush] [is proposing] [that] [workers] [be allowed] [to divert] [their payroll taxes] [into private accounts] . ##

Figure A.16: This alignment is marked as Genitive + Coreference + Reasoning, since there is additional reasoning involved, related to knowledge about proper names and titles.



--728-YES-SUM--

With stars like Sluggers Barry Bonds under suspicion and lawmakers demanding action , [Major League Baseball] [adopted] [a tougher steroid-testing program] [that will suspend] [first-time offenders] for 10 days [and] [randomly] [test] players [year-round] . #

[Baseball] [rolled out] [a tougher steroid-testing program] [that includes penalties] [for first-time offenders] [and] [random] [, year-round] [checks] . ##

Figure A.17: The aligned markables *random* and *randomly* are both in modifier positions and their heads (*checks* and *test*, respectively) are aligned. Therefore their alignment is not labeled as Modifier but simply as Morphological of type Other.

4. Finally, Principle 4. regarding additional labels for the case of Genitive equally applies to the case of Modifier.

Morphological. The definition of this feature is straightforward and does not require many additional explanations beyond the ones provided in Section 3.1.1.

Nominalization. We note that in case the verb-markable is in passive voice, then the Passivization feature is also selected, as in Figure A.19.

Demonym. Straightforward.

Acronym. Straightforward.

Other. Similarly to Nominalization, this feature (and not Identity) is used for the rare case of two identical words aligning, but appearing as different parts of speech; e.g., *year-round* as adjective and as adverb.

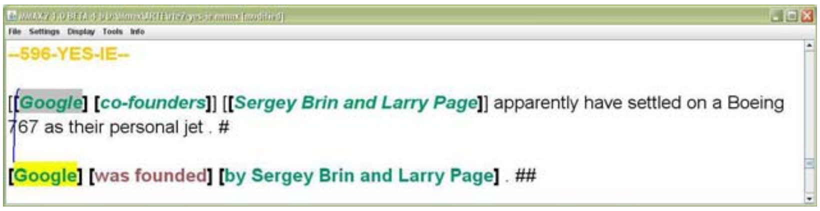
Finally, it also applies to complex morphological transformations involving more than one of the above processes; e.g. *TV* ⟷ *televised*, which in fact implicates both nominalization and acronym transformations.

Argument Variation. The use of this feature takes into account also **oblique** arguments (i.e., arguments which are not subjects or objects—in English always marked with prepositions).

In case the predicates aligned describe the same prototypical event from two different perspectives, as in Figure 3.7, then the alignment takes a positive Argument Variation, but a negative Reasoning value. In contrast, if there is argument variation between two aligned predicates which are not generally considered as describing exactly the same event, as in Figure A.20, then Reasoning is additionally marked.

It should nonetheless be emphasized that the feature applies only if both markables aligned are verbs. This is not the case in Figure A.21, for example.

Passivization. Necessary condition for the use of this feature is a *straightforward* Argument Variation alignment—in fact, Passivization is a dependent feature of Argument Variation. For example, in Figure A.20 this was not the case, since the two predicates are linked by Reasoning. However in Figure A.22 the predicates are linked by Synonymy and hence Passivization applies.

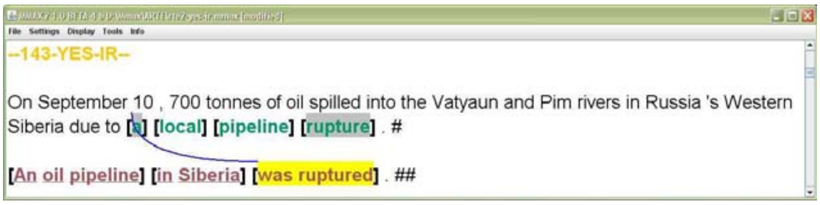


--596-YES-IE--

[[Google] [co-founders]] [[Sergey Brin and Larry Page]] apparently have settled on a Boeing 767 as their personal jet . #

[Google] [was founded] [by Sergey Brin and Larry Page] . ##

Figure A.18: The H-markable *was founded* is in a straightforward (labeled as Nominalization + Passivization + Hypernymy) alignment to the T-markable *co-founders*. Hence its modifier, *Google*, which simply aligns to the subject of the predicate *was founded*, will not be associated with the Modifier, but with the Identity feature.

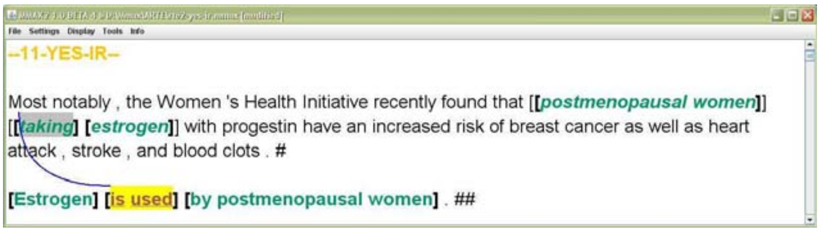


--143-YES-IR--

On September 10 , 700 tonnes of oil spilled into the Vatyaun and Pim rivers in Russia 's Western Siberia due to [a] [local] [pipeline] [rupture] . #

[An oil pipeline] [in Siberia] [was ruptured] . ##

Figure A.19: An alignment marked as Nominalization + Passivization, since the predicate appears in passive voice.



--11-YES-IR--

Most notably , the Women 's Health Initiative recently found that [[postmenopausal women]] [[taking] [estrogen]] with progestin have an increased risk of breast cancer as well as heart attack , stroke , and blood clots . #

[Estrogen] [is used] [by postmenopausal women] . ##

Figure A.20: An alignment marked as Argument Variation + Reasoning, since the predicates *take* and *use* are not considered as describing the same prototypical action.

Ontological. The definition of synonymy can be a cause of disagreement for theorists, as well as annotators.

Synonymy. It directly implies that the markables are of the same syntactic category, as expressions of different syntactic categories cannot be interchanged without making the sentence ungrammatical.

However the definition of synonymy[6] is by nature quite elusive, since, as most linguists and psychologists argue, it is not a discrete but rather a gradient concept. In cases of uncertainty the alignment is not labeled as Synonymy, but as Reasoning.

Hypernymy. Straightforward.

Quantities. If an NP includes a quantifier, then the Alignment-markable constructed from it must include every single part of the NP constituent, like modifiers, since theses are indispensable for a correct semantic interpretation.

Note also that the notion of *scalar implicatures* is highly relevant in this case, and should be considered when making entailment judgments.

For instance, as Levinson (1983) notes, the quantifiers *none*, *some* and *all* constitute an **implicational scale**, i.e., a list of lexical items of the same constituent category that are ordered in terms of their informativeness. The ordered list <*always*, *often*, *sometimes*> is another example. In such a scale the use of one form implicates that the use of a stronger form is not possible.

Reasoning. In more detail the different forms of reasoning covered by this feature include:

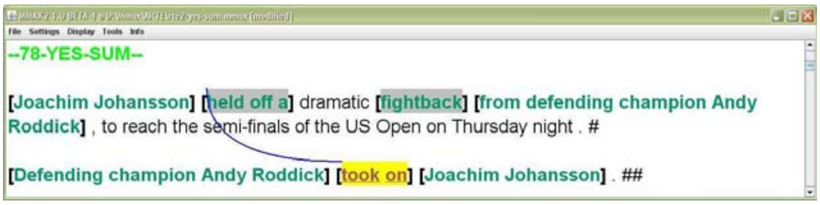
1. A **lexical relation** not among the ones listed above. That could be
   - a functional relation like holonymy, is-made-of, is-an-attribute-of (e.g., *Congress* ⟷ *the government*, *troops* ⟷ *five soldiers*);

[6] (Miller et al., 1990) contains an interesting discussion on the concept of **synonymy**. As Miller et al. remark, the traditional definition, which demands interchangeability of the two expressions in all possible contexts, is rather too restrictive and makes true synonymy questionable. For this reason the weakened version, which examines synonymy with respect to context, is considered preferable.

  - a more complex relation such as lexical entailment, enablement, cause or happens-before (e.g., *was sold* ⟷ *cost*, *arrived* ⟷ *is visiting*);
  - an idiom or a paraphrase (e.g., *saw the light of day* ⟷ *was released*);

2. **General world knowledge**, such as knowledge of entities and the relationships among them, e.g., *NASA* ⟷ *U.S.*, *David Hiddleston* ⟷ *a person*;
3. **Geographical knowledge** in particular; e.g., *Gaza* ⟷ *Gaza Strip*;
4. **Spatial knowledge**; e.g., *in* ⟷ *is located in*;
5. **Temporal knowledge**, as in Figure A.23;
6. Interpretation of **modality** information, carried by modal auxiliary verbs such as *may*, *must*, *can* (e.g., *may not be safe* ⟷ *are not safe*).

   We note that in this case the task definition regarding the judgment of probable—but not certain—inferences, as presented in Section 1.1, is highly relevant and should be taken into account.
7. Syntactic structures whose interpretation is strongly influenced by **punctuation** markers, as was presented in Figure A.3;
8. **Logical inference** mechanisms, as in Figure A.24;
9. Other **general inference** mechanisms, as in Figure A.25;
10. Interpretation of figures of speech such as **metonymy** (i.e., the trope in which one entity is used to stand for another, closely associated, entity), or the more specific **synechdoche** (i.e., the trope in which a part or constituent stands for a whole or a more comprehensive entity it belongs to) and **antonomasia** (i.e., the substitution of an epithet, description or name with a related proper name). An example was provided by Figure 3.9;
11. Interpretation of an **elliptical construction** (i.e., a construction that lacks elements recoverable from the context), as was the case in Figure 3.10;
12. Interpretation of a **conversational implicature**, as in Figure A.26;
13. Interpretation of the sentence's **context**, in case the Coreference feature does not apply, i.e., the aligned pair is not a pair of NP-markables. An example of this is provided by Figure A.27.

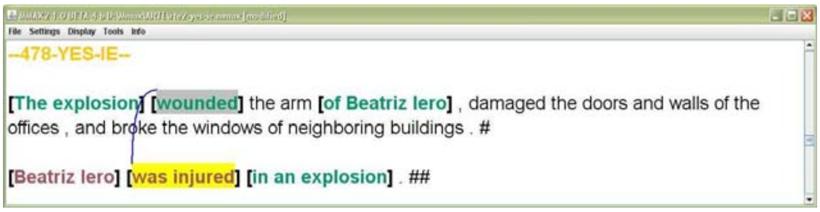


--78-YES-SUM--

[Joachim Johansson] [held off a] dramatic [fightback] [from defending champion Andy Roddick] , to reach the semi-finals of the US Open on Thursday night . #

[Defending champion Andy Roddick] [took on] [Joachim Johansson] . ##

Figure A.21: The T-markable *held off a fightback* is not a verb but a verb phrase, and hence the alignment is labeled solely as Reasoning, and not as Argument Variation.

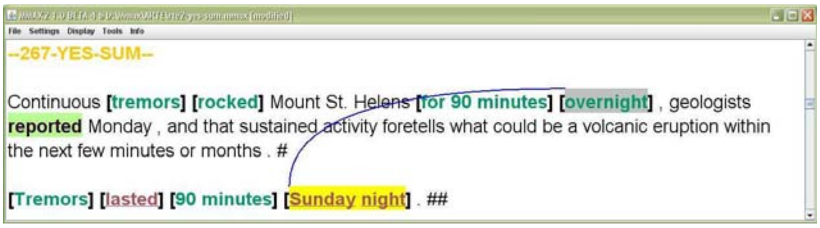


--478-YES-IE--

[The explosion] [wounded] the arm [of Beatriz Iero] , damaged the doors and walls of the offices , and broke the windows of neighboring buildings . #

[Beatriz Iero] [was injured] [in an explosion] . ##

Figure A.22: A Passivization alignment, even though the markables are not forms of the same predicate.

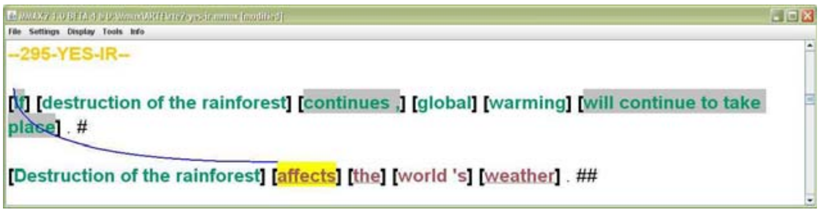


--267-YES-SUM--

Continuous [tremors] [rocked] Mount St. Helens [for 90 minutes] [overnight] , geologists reported Monday , and that sustained activity foretells what could be a volcanic eruption within the next few minutes or months . #

[Tremors] [lasted] [90 minutes] [Sunday night] . ##

Figure A.23: This alignment, marked as Reasoning, demands complex reasoning based on knowledge about the temporal precedence of the days of the week.

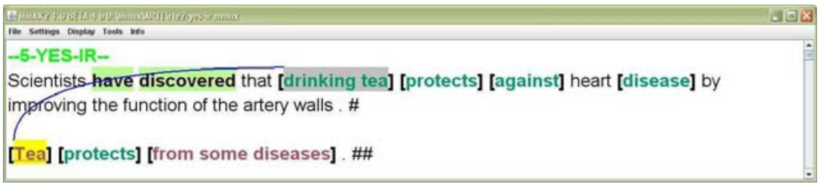


--295-YES-IR--

[If] [destruction of the rainforest] [continues ,] [global] [warming] [will continue to take place] . #

[Destruction of the rainforest] [affects] [the] [world 's] [weather] . ##

Figure A.24: An alignment marked as Reasoning since it is based on the interpretation of the conditional logical connective.

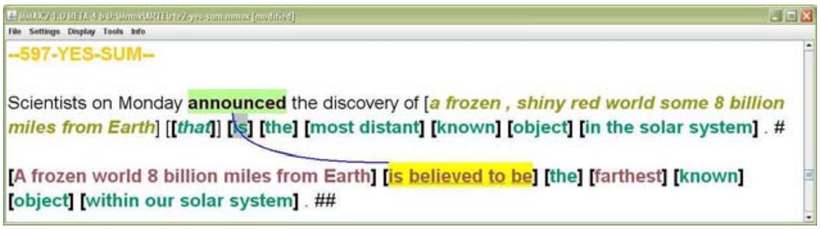


--5-YES-IR--

Scientists have discovered that [drinking tea] [protects] [against] heart [disease] by improving the function of the artery walls . #

[Tea] [protects] [from some diseases] . ##

Figure A.25: An alignment marked as Reasoning since it is based on general inference procedures.

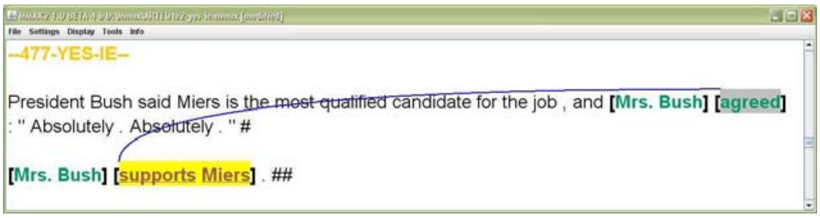


--597-YES-SUM--

Scientists on Monday announced the discovery of [*a frozen , shiny red world some 8 billion miles from Earth*] [[*that*]] [is] [the] [most distant] [known] [object] [in the solar system] . #

[A frozen world 8 billion miles from Earth] [is believed to be] [the] [farthest] [known] [object] [within our solar system] . ##

Figure A.26: The alignment relies on a conversational implicature raised by the Maxim of Quality, according to Grice (1975). Thus it is labeled as Reasoning.



--477-YES-IE--

President Bush said Miers is the most qualified candidate for the job , and [Mrs. Bush] [agreed] : " Absolutely . Absolutely . " #

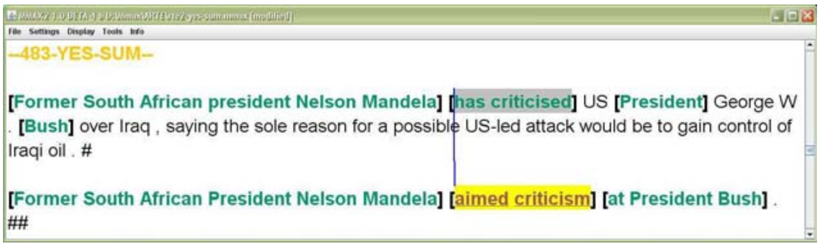
[Mrs. Bush] [supports Miers] . ##

Figure A.27: This alignment is heavily based on the interpretation of the context and is marked with a positive Reasoning value.

A.2.3 Multilabeled Alignments

It should be emphasized that, as must have already become evident by some of the examples (e.g., Figures A.19 and A.20), the features presented above are not mutually exclusive.[7] Therefore we can—and must—in certain cases use them in combination, in order to indicate all the grammatical phenomena present and describe the alignment in the best possible way.

There is no restriction posed by the annotation scheme regarding how many different labels an alignment may have—the upper bound is the number of different features available, as long of course, as their combination is sensible and purposeful. In other words, if the inference responsible for the alignment has already been captured by the selection of one feature, the alignment will not additionally be labeled with extra features, even if they are applicable. Figure A.28 illustrates this.



--483-YES-SUM--

[Former South African president Nelson Mandela] [has criticised] US [President] George W . [Bush] over Iraq , saying the sole reason for a possible US-led attack would be to gain control of Iraqi oil . #

[Former South African President Nelson Mandela] [aimed criticism] [at President Bush] .
##

Figure A.28: This alignment involves a paraphrase and is thus labeled solely with Reasoning; the Nominalization feature would be redundant.

The practice of multilabeling naturally applies also to contextual alignments[8]—either of the Coreference or the Reasoning type. For any contextual alignment there are two cases to consider:

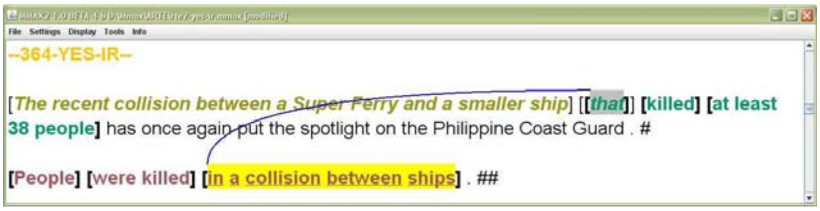
Case 1: T-markable linked to its context at the Coreference level. In this case the contextual alignment will additionally be labeled with any features required to describe the transition/inference from the relevant part of T (i.e., the context) to the H-markable. An example is provided by Figure A.29.

Case 2: The Coreference level does not indicate the context. Here we will not aim at a detailed description of the type of inference involved, since no particular part of T has been pointed out as respon-

[7] Unlike the different values defined for each feature, which are.

[8] For a detailed discussion of contextual alignments see Subsection A.4.3

sible for the contextual alignment. Instead we will simply apply the additional label of Reasoning, as in Figures A.9 and A.27.



--364-YES-IR--

[*The recent collision between a Super Ferry and a smaller ship*] [[*that*]] [killed] [at least 38 people] has once again put the spotlight on the Philippine Coast Guard . #

[People] [were killed] [in a collision between ships] . ##

Figure A.29: This alignment is marked as Hypernymy + Reasoning in addition to Coreference, since the inference is based on the knowledge that a *Ferry* is a kind of *ship*, and on general reasoning. Of course the H-markable *that* is linked to its antecedent through a Coreference pointer relation.

A.3 Context

The Context level provides information related to factivity and negative polarity factors.

A.3.1 Markables

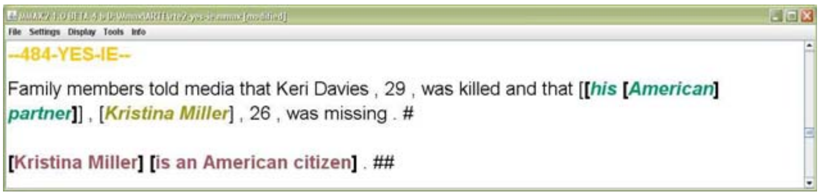
Annotated are embedding predicates and other kinds of complement-taking constructions or expressions that may affect the semantic behavior of the sentence they introduce. The selection of markables at this level is guided by two principles.

First, the relative polarity imposed by the expression must be important for the entailment value of the pair; otherwise it will not be marked. Figure A.30 presents such a case. Second, the semantic contribution of the expression must not have already been captured at the Alignment level, as for example in Figure A.31.

Therefore it is clear that not every T–H pair is annotated at this level; Context annotation is created rather when there is important context information which cannot be incorporated at the Alignment level by constructing direct mappings between T and H.

A.3.2 Features

We investigate two different features, both of the attribute type, as presented in Figure A.32.

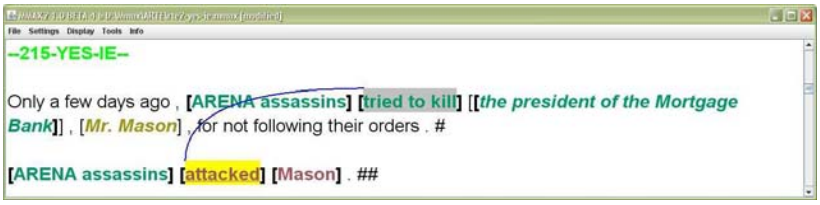


--484-YES-IE--

Family members told media that Keri Davies , 29 , was killed and that [[his [American] partner]] , [Kristina Miller] , 26 , was missing . #

[Kristina Miller] [is an American citizen] . ##

Figure A.30: This pair is similar to the one of Figure 3.11 in that the that-complement clause in T is in the context of a neutral factivity expression (*told*). However the entailment here is entirely based on presuppositions triggered by the possessive phrase *his American partner* and the apposition construction. Therefore the factivity of *told* is irrelevant in this case, and no Context level annotation is created.

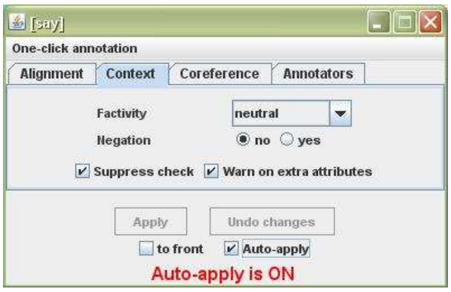


--215-YES-IE--

Only a few days ago , [ARENA assassins] [tried to kill] [[the president of the Mortgage Bank]] , [Mr. Mason] , for not following their orders . #

[ARENA assassins] [attacked] [Mason] . ##

Figure A.31: The accomplishment verb *tried* in T introduces a particular factivity context—namely, it is a one-way –implicative. However it is not marked at the Context level, since its semantics are captured by a direct alignment.



Factivity neutral

Negation ● no ○ yes

☑ Suppress check ☑ Warn on extra attributes

□ to front ☑ Auto-apply

Auto-apply is ON

Figure A.32: The features that apply to the Context markable *say* of Figure 3.11.

Factivity. The use of this feature is mostly self-explanatory. We make a few additional comments, especially with regard to the case of implicative expressions.

Neutral. Straightforward.

Factive. Straightforward.

Counterfactive. Straightforward.

Implicative. As Nairn et al. (2006) point out, in certain cases it is hard to distinguish *entailments* from *conversational implicatures.* For example the expression *be able* in a positive relative polarity induces a conversational implicature, which however may be contradicted by further context. On our annotation we include such expressions in this category.

It should also be noted that many typically implicative predicates such as *manage*, *refuse* and *attempt* belong in this category only when they introduce nonfinite complements (e.g. infinitives or participles). The very same predicates may play a radically different role if their complement clause is finite. For example let us consider again the sentences in (2) of Subsection 3.1.2.

(2) a. Ed forgot that he had closed the door.
b. Ed did not forget that he had closed the door.
c. Ed forgot to close the door.
d. Ed did not forget to close the door.

The predicate *forget* acts as factive when it introduces a finite that-complement as in (2a) and (2b), since both sentences presuppose the complement's truth. On the other hand it functions as a two-way implicative in the sentences with nonfinite complements (2c) and (2d), as (2c) implies the falsity of the complement but its negation (2d) implies the complement's truth.

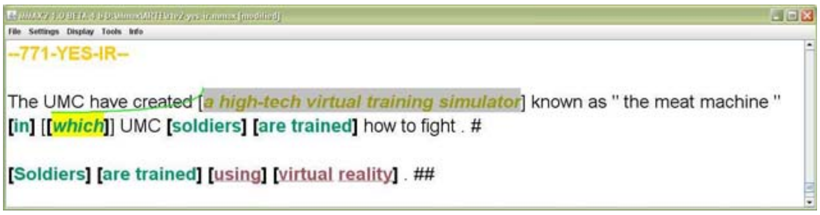
Negation. Straightforward.

A.4 Coreference

While the Coreference feature of the Alignment level labels cross-sentential coreference between T and H, the Coreference level captures coreference within the boundaries of T, which is usually intra-sentential.

A.4.1 Markables

At the Coreference level the markables are formed by the complete NPs[9], including any modifiers and complements of the head nouns. However, as Figure A.33 illustrates, they do not include relative clauses subordinate to them, or appositive NPs. Such relative clauses are annotated as separate markables in case they carry information significant for the entailment, as Figure A.34 exemplifies.

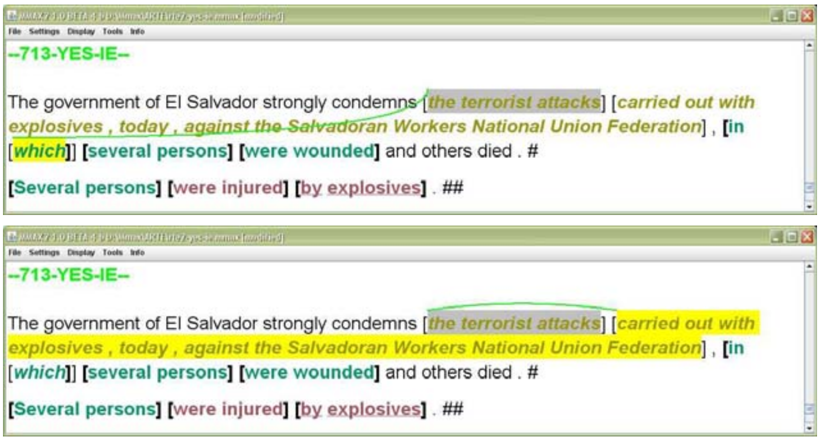


--771-YES-IR--

The UMC have created [a high-tech virtual training simulator] known as " the meat machine " [in] [[which]] UMC [soldiers] [are trained] how to fight . #

[Soldiers] [are trained] [using] [virtual reality] . ##

Figure A.33: The reduced relative clause *known as "the meat machine"*, which modifies the head noun *simulator* is not part of the markable.


File Settings Display Tools Info

--713-YES-IE--

The government of El Salvador strongly condemns [the terrorist attacks] [carried out with explosives , today , against the Salvadoran Workers National Union Federation] , [in [which]] [several persons] [were wounded] and others died . #

[Several persons] [were injured] [by explosives] . ##

File Settings Display Tools Info

--713-YES-IE--

The government of El Salvador strongly condemns [the terrorist attacks] [carried out with explosives , today , against the Salvadoran Workers National Union Federation] , [in [which]] [several persons] [were wounded] and others died . #

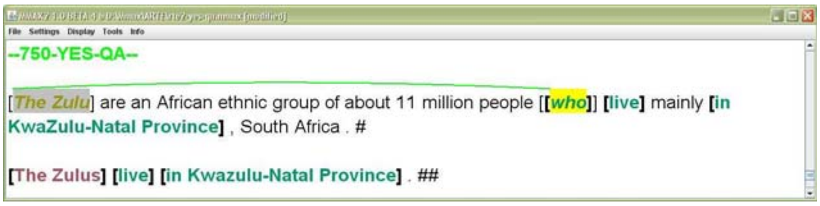
[Several persons] [were injured] [by explosives] . ##

Figure A.34: The reduced relative clause *carried out . . . Federation* does not form part of the NP-markable *the terrorist attacks*, but since it provides crucial information for establishing the entailment, it is linked to it as an individual markable.

[9] In case the markables in question are NP-markables. Of course other types of markables are possible, such as nouns, pronouns and reduced relative clauses.

In any case, Coreference markables are obviously only created when the coreference relation plays some role for the entailment, i.e., carries information that is important for one of the alignments. The Coreference level is in this sense subsidiary to the Alignment level. Figure A.35 offers an example of such a selective annotation.

Additionally, Figure A.36 depicts a pair in which Coreference annotation has been spared. For simplicity, Coreference annotation is created only in T, and any coreferences in H are ignored, as illustrated in Figure A.10.

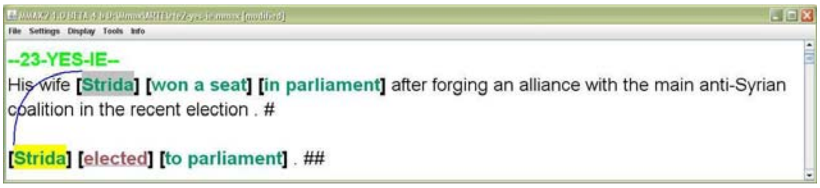


--750-YES-QA--

[The Zulu] are an African ethnic group of about 11 million people [[who]] [live] mainly [in KwaZulu-Natal Province] , South Africa . #

[The Zulus] [live] [in Kwazulu-Natal Province] . ##

Figure A.35: In this pair the Coreference annotation purposefully ignores the NP *an African ethnic group of about 11 million people.*



--23-YES-IE--

His wife [Strida] [won a seat] [in parliament] after forging an alliance with the main anti-Syrian coalition in the recent election . #

[Strida] [elected] [to parliament] . ##

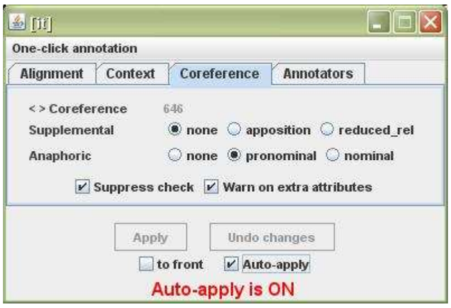
Figure A.36: Here the two NPs of T *his wife* and *Strida* are not linked at the Coreference level, since their coreference is not relevant for the entailment.

A.4.2 Features

Three different features have been defined at this level, as depicted in Figure A.37.

In the first place, coreferent NPs in T are linked by means of the Coreference pointer relation, whose direction is determined by examining the specific type of the coreference in question and is elaborated below. This relation is very similar to Alignment, in the respect that it creates ordered pairs of markables, though, unlike Alignment, both markables are from T.

Also like Alignment, Coreference is branching and, once built, it reveals an additional set of features that function as labels for its type:



< > Coreference 646

Supplemental ● none ○ apposition ○ reduced_rel

Anaphoric ○ none ● pronominal ○ nominal

☑ Suppress check ☑ Warn on extra attributes

□ to front ☑ Auto-apply

Auto-apply is ON

Figure A.37: The features applying to the Coreference-markable *it* of Figure 3.12.

Supplemental. There are several principles involving the modeling of each type of supplemental expression.

Apposition. As in paragraph A.2.1, we differentiate between two distinct cases that call for different treatment.

Case 1: Punctuation marks the appositive. Then the direction of the pointer relation is from the NP marked by punctuation (i.e., the appositive) to the main NP.

Case 2: The appositive is not marked by punctuation. Then the direction of the pointer relation is determined by the way the Alignment-markables have been selected. In case only one of the two Coreference-markables is involved or partly involved in an alignment, then the direction of the Coreference relation is from the one which is not active at the Alignment level to the one that is.
Otherwise, in case both of the Coreference-markables are to some extent also active at the Alignment level, then the alignment guides us as to which one of the two is more **prominent** than the other in our case, in the following sense:

Normally one of the two Coreference-markables will be aligned to a coreferent NP in H—usually the subject—, while the other one will not be involved in any alignment, or will align to a predicate or object in H. Then the latter will be regarded as the less prominent and will be chosen as the source of the Coreference relation, with the former as the target. Figure A.38 illustrates this.

Reduced Relative. The relation points from the reduced relative clause to the NP, as shown in Figure 3.14.

Our definition of a **reduced relative clause** involves a relative clause with a nonfinite verb (i.e., in this case, a participle), which lacks a relative pronoun or relativizer. This definition requires that the reduced relative clause is a VP and, unlike the analysis of Sag (1997), admits the clauses in (4) but not the ones in (5).

(4) a. the person [standing on my foot] ...
b. the prophet [descended from heaven] ...
c. the bills [passed by the House yesterday] ...

(5) a. the people [in Rome] ...
b. the people [happy with the proposal] ...

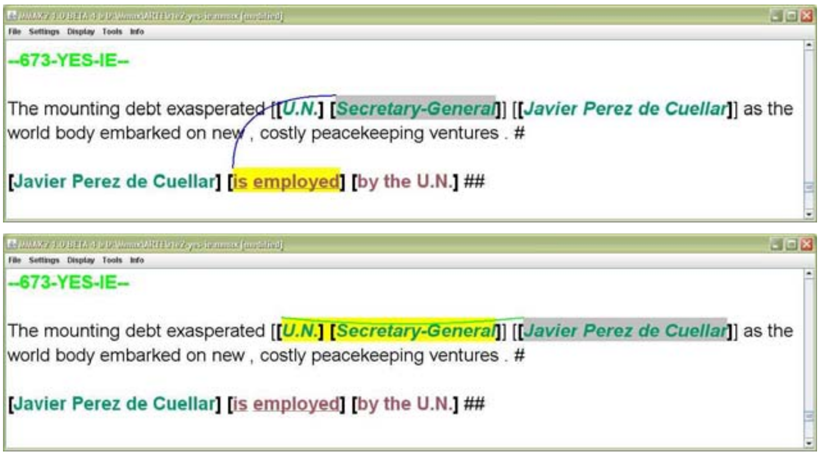
Anaphoric. Fairly straightforward; we merely make a few remarks:

Pronominal. The feature applies also to anaphora induced by the relativizer *that*. The direction of the Coreference relation is from the pronoun/relativizer to the antecedent.

Nominal. An **equative** clause is one in which the main predicate is the copula and the semantic predicate is expressed by an NP, like *George Washington was the first President of the United States.* Figure A.39 presents such an example, which is not annotated at the Coreference level.

The direction of the Coreference pointer relation is from the NP that is involved in a Coreference Alignment, to the context NP of T.

Finally, we note that the Anaphoric coreference feature is applied also in cases where the markables are not strictly coreferent, but one of them is a member of the class named by the other, as in Figure A.40.



--673-YES-IE--

The mounting debt exasperated [[U.N.] [Secretary-General]] [[Javier Perez de Cuellar]] as the world body embarked on new , costly peacekeeping ventures . #

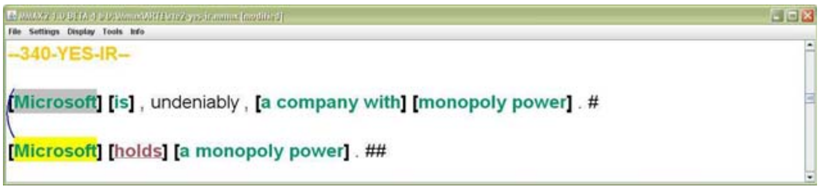
[Javier Perez de Cuellar] [is employed] [by the U.N.] ##

--673-YES-IE--

The mounting debt exasperated [[U.N.] [Secretary-General]] [[Javier Perez de Cuellar]] as the world body embarked on new , costly peacekeeping ventures . #

[Javier Perez de Cuellar] [is employed] [by the U.N.] ##

Figure A.38: The Coreference-markable *Javier Perez de Cuellar* is involved in an Identity alignment with the subject of H, while the head of the Coreference-markable *U.N. Secretary-General* is aligned to the predicate of H. Therefore the direction of the Coreference relation is from the latter to the former, which is seen as the most prominent of the two.

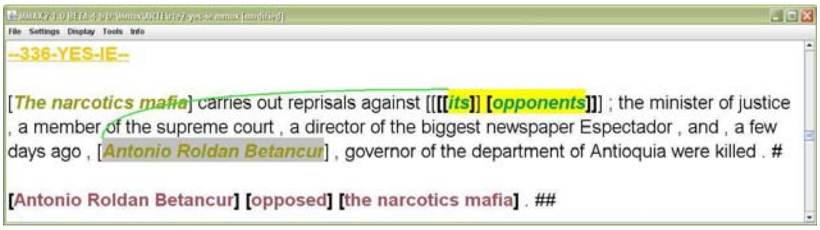


--340-YES-IR--

[Microsoft] [is] , undeniably , [a company with] [monopoly power] . #

[Microsoft] [holds] [a monopoly power] . ##

Figure A.39: In T the predicate complement *a company with monopoly power* is in an equative clause and hence will not be linked to the NP *Microsoft* by a Coreference relation.



--336-YES-IE--

[The narcotics mafia] carries out reprisals against [[[its]] [opponents]]] ; the minister of justice , a member of the supreme court , a director of the biggest newspaper Espectador , and , a few days ago , [Antonio Roldan Betancur] , governor of the department of Antioquia were killed . #

[Antonio Roldan Betancur] [opposed] [the narcotics mafia] . ##

Figure A.40: The markable *Antonio Roldan Betancur* is directly linked by Anaphoric coreference to the markable *its opponents*, since it denotes a member belonging to this class.

A.4.3 Another Look at Contextual Alignments

At this point the presentation of the basic levels of the annotation is complete. Now that we have the whole picture, we can reflect back on the most fundamental part of the scheme, the Alignment level, and shed some more light on certain less transparent aspects of it. In particular, a little more detail is in order about how we produce the types of alignment we call **contextual**, i.e., not based on direct relations and inferences between two markables, but rather induced by the wider interpretation of the T-markable, drawn by the context in which it appears.

According to our definitions, there are two different cases of contextual alignment implemented in the annotation scheme. The first case is when the two markables aligned are coreferent NPs. This type of alignment is extensively discussed on the definition of the Coreference feature in Subsection A.2.2. The second is when the alignment is between two non-NP markables, as illustrated in Figure A.27. In this case the markables are on the one hand obviously related semantically, due to the interpretation forced to the T-markable by its context; on the other hand there is no coreference involved and the alignment is marked as Reasoning.

The indispensability of the contextual alignments is evident, considering the fact that our alignments are heavily driven by the syntactic structures of T and H, which they strive to respect as much as possible. According to this principle, once a predicate in H is aligned to a T-markable, then the arguments and adjuncts of this predicate will be attempted to get aligned to markables of T in direct syntactic relations to the initial T-markable. This is for example the reason behind the contextual alignments of the two subject-markables of H in Figures A.9 and 3.3, although the T-markables are on their own insufficient for establishing the entailments.

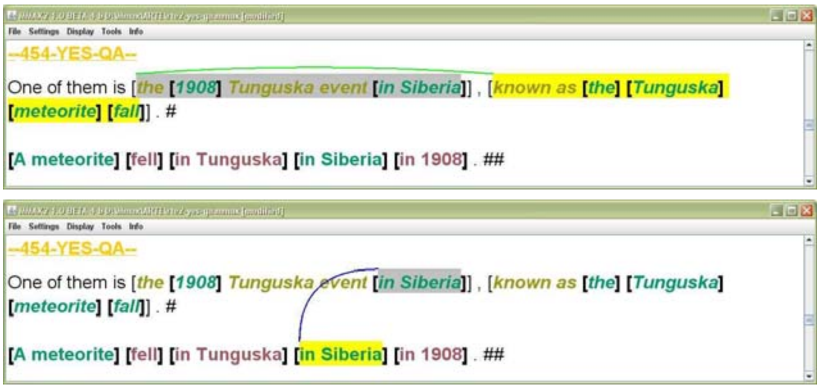
On the other hand, however, sometimes we violate the constraints posed by syntactic structure and proceed to direct alignment of constituents which would traditionally only participate in contextual alignments, as Figure A.41 exemplifies.

The question arising is natural: how do we determine in which cases a markable of H should participate in a direct alignment to a corresponding part in T, even violating the syntactic structure, and when it should be aligned contextually, respecting the local dependencies? The key to the answer is given, as one would suspect, by the Coreference level, which, as we have seen so far, supplements the level of Alignment in a variety of ways. Namely we adopt the following principles:

Case 1. If T has not been marked at the Coreference level, and therefore there is no pointer to the context responsible for the alignment, as was the case in Figure A.27, then a contextual alignment is made—labeled either as Coreference or as Reasoning, as explained above.

Case 2. If the T-markable has been linked to some other part of T through a Coreference relation of type Anaphoric, then again it will be involved in a contextual alignment, as in Figure A.42.

Case 3. Finally, if we have a Coreference relation of type Supplemental—either Apposition or Reduced Relative—between the part of T which is a candidate for participating in a contextual alignment, and its context in T, which establishes the local entailment, then we are "entitled" to disobey the syntactic structure rules, and make a direct alignment between the relevant parts. Exactly this was the case in Figure A.41.


File Settings Display Tools Info
File Settings Display Tools Info

--454-YES-QA--

One of them is [the [1908] Tunguska event [in Siberia]] , [known as [the] [Tunguska] [meteorite] [fall]] . #

[A meteorite] [fell] [in Tunguska] [in Siberia] [in 1908] . ##

--454-YES-QA--

One of them is [the [1908] Tunguska event [in Siberia]] , [known as [the] [Tunguska] [meteorite] [fall]] . #

[A meteorite] [fell] [in Tunguska] [in Siberia] [in 1908] . ##

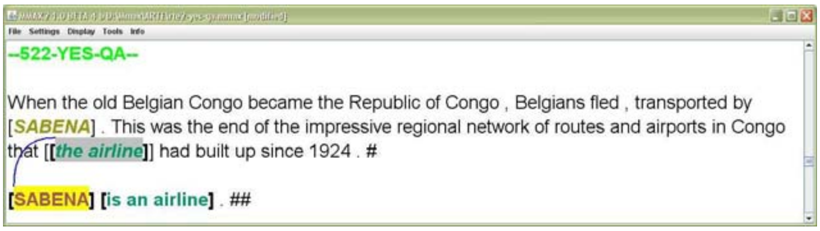
Figure A.41: In this pair the main predicate in H, *fell*, aligns to the markable *the fall* of the reduced relative clause in T. However its argument, *in Siberia*, does not get directly aligned to some part of the reduced relative clause, and violates the syntactic structure through an alignment to a part of the main clause of T.

A.5 Non-entailment

The Non-Entailment level is used to model the negative entailment pairs.

A.5.1 Markables

The selection of markables is rather minimalist, compared to the one for the positive pairs. We do not exhaustively include the whole H in markables; in fact it is possible that either T or H (but not both) does not contain any



--522-YES-QA--

When the old Belgian Congo became the Republic of Congo , Belgians fled , transported by [SABENA] . This was the end of the impressive regional network of routes and airports in Congo that [[the airline]] had built up since 1924 . #

[SABENA] [is an airline] . ##

Figure A.42: The information expressed in H is based on a Nominal coreference between the T-markables *the airline* and *SABENA*. Therefore the alignment of the H-markable *SABENA* will be contextual.

markables at all. On top of that, as the annotation aims at pointing out the one most important factor that blocks the entailment, maximally one markable in each of the T and H of each pair is created.

The markables are created in such a way that they generally correspond to constituents; e.g., V-, NP-, AdjP-, Adv- and PP-markables. For the selection of markables predicates or copula and predicate complement constructions are generally given priority; subjects and objects are then considered, in the order mentioned. The NP markables are formed following the same[10] guidelines as at the Alignment level before, described in Section A.2.1. Again, however, the main consideration is semantic interpretation and not syntactic constituency.

A.5.2 Features

As the annotation of the negative entailment pairs serves the purpose of simplicity and straightforwardness rather than detail, the annotation scheme here is much more plain than in the positive entailment case. It contains one basic branching feature of the attribute type, Non-entailment. An appropriate attribute selection on this feature unfolds two dependent features of the pointer type, as presented in Figure A.43.

Non-entailment can take one of the following available values:

Context. This value is selected in case the remaining contents of T and H, excluding the context, could possibly be considered as in an entailment relation, and it is mainly due to the interference of the context that we are not allowed to establish it.

[10] With the exception of the requirements for the alignment of appositives, which are dropped.

Additional. Clearly only parts of H are marked as Additional. What is more, such an annotation indicates that the additional information is unrelated enough with respect to T, to be incapable of participating in a direct alignment to some part of it.

Note that the sense of *alignment* here is different than the one used for the positive entailment cases. What is meant in this case is rather a *misalignment*, as elaborated on the definition of the corresponding value for Nonentailment.

Misalignment. This value differs from the value Additional in that here T and H are highly related from a semantic point of view, and possibly even have similar syntactic structures.

Once it is selected, it brings out the following two dependent features, which serve as specifying labels of the Misalignment, much in the same way as the dependent features of Alignment do in the positive scheme. Both of them are pointer relations directed from the H-markable to the corresponding T-markable that is responsible for the misalignment.

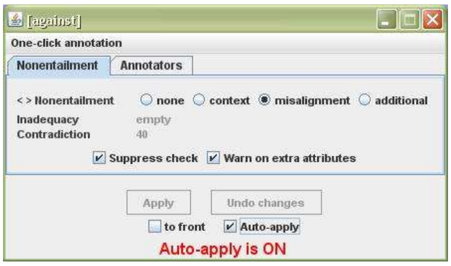
Inadequacy. Straightforward.

Contradiction. The notion of contradiction can be obscure when regarded outside a strictly defined axiomatic system. In particular when it comes to natural language sentences describing events/situations in the world, contradiction becomes relative to whether the two statements are interpreted in exactly the same spatiotemporal frame, referring to the same events/situations, or not.

For instance, in Figure 3.16, if the misaligned pair is not viewed in the same event-specific frame evoked by its context, it does not induce a contradiction; e.g. one could well accept the truth of both sentences at different points in time.

However, following (Manning et al., 2007), we assume that in absence of countervailing evidence:

1. Compatible NPs between T and H are coreferent;
2. Apparently overlapping descriptions of events/situations in T and H refer to the same spatiotemporal frame.



< > Nonentailment ○ none ○ context ● misalignment ○ additional

Inadequacy empty

Contradiction 40

☑ Suppress check ☑ Warn on extra attributes

□ to front ☑ Auto-apply

Auto-apply is ON

Figure A.43: The attribute window, containing the set of features for the selected markable of Figure 3.16. It indicates a Contradiction.

A.6 Annotators

As an extra level of annotation, we designed one especially dedicated to monitoring the annotators' impressions during their work. In this sense, the Annotators level is not part of the actual annotation, but may rather be considered as a meta-level.

A.6.1 Markables

The purpose of this level is to record certain aspects of the annotators' attitude towards each individual pair of the datasets. Therefore the markables here will be the ID codes of the pairs. All pairs must be marked, resulting in a number of Annotators-markables equal to the number of pairs.

A.6.2 Features

There are two features at this level, as seen in Figure A.44.

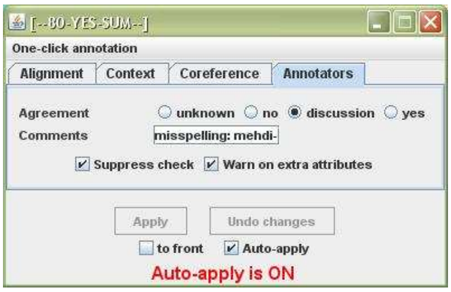
Agreement. It is specified during the second phase of the annotation, in which the annotators compare and discuss their individually created annotations. It indicates the level of inter-annotator agreement reached after this process. It takes one of the following values:

- **No.** It signals that the annotators have radically different views about the most appropriate annotation of the pair, and no agreement can be reached.
- **Discussion.** It indicates that agreement between the annotators is reached, but only after discussion clearing original points of conflict.

This does not relate to annotation errors, but to a linguistic discussion about the best possible modeling of the data. In this sense, it could provide an informal indication of pairs requiring a deeper analysis.

Yes. It indicates that after correction of possible errors the annotators support exactly the same annotation. Contrary to Discussion, this value provides indication of possibly dealing with a simple-to-analyze pair.

Comments. This is a free-text field, originally empty. It serves mainly the annotators' convenience and its use is left entirely to their judgment.



[--80-YES-SUM--]

One-click annotation

Agreement ○ unknown ○ no ● discussion ○ yes

Comments misspelling: mehdi-

☑ Suppress check ☑ Warn on extra attributes

□ to front ☑ Auto-apply

Auto-apply is ON

Figure A.44: The attribute window for the Annotators-markable –*80-YES-SUM*–.