Comparing methods for the syntactic simplification of sentences in information extraction

Richard J. Evans
University of Wolverhampton, UK

Abstract

This article describes research aimed at improving the accuracy of an information extraction (IE) system by treating coordinate structures systematically. Commas, coordinating conjunctions, and adjacent comma–conjunction pairs are considered to be potential indicators of coordination in natural language. A recursive algorithm is implemented which converts sentences containing classified potential coordinators into sequences of simple sentences. Several approaches to the classification of potential coordinators are presented, one exploiting memory-based learning, another exploiting the publicly available Stanford parser, and a hybrid approach that classifies commas and conjunctions using the former system and comma–conjunction pairs using the latter. The article describes the initial set of features developed for exploitation by the memory-based classifier and presents optimization of that classifier. A baseline system is also described. The sentence simplification module was exploited by an IE system. With regard to the automatic classifiers that form the basis for simplification, comparative evaluation demonstrated that IE can be performed with greatest accuracy when exploiting the hybrid classifier. It also demonstrated that a simple baseline classifier induces improved accuracy when compared to systems that ignore the presence of coordinate structures in input sentences. The article presents an analysis of the errors made by the different sentence simplification modules and the IE system that exploits them. Directions for future research are suggested.

Correspondence:
Research Institute in Information and Language Processing,
University of Wolverhampton,
Stafford Street,
Wolverhampton,
West Midlands,
WV1 1SB, UK.
E-mail:
R.J.Evans@wlv.ac.uk

1 Introduction

This article presents a method to improve the accuracy of a clinical information extraction (IE) system by pre-processing syntactically complex input sentences. The research described investigates the automatic simplification of syntactic complexity in natural language. It focuses on the relations of ‘subordination’ and ‘coordination’ that involve the linking of syntactic units of the same rank in a sentence. In subordination, the linked units form a hierarchy with the subordinate unit being a constituent of the superordinate unit (1).

(1) [[For the past 3 days][,] he has had fever, malaise, and headache].

In coordination, the linked units are constituents at the same level of constituent structure (2). It is ‘a type of linkage whereby the resulting conjoint construction is equivalent, structurally speaking, to each of its members’. That is, ‘if [A] and [B] are conjoins of the conjoint construction X, then any structural

doi:10.1093/llc/fqr034 Advance Access published on 31 August 2011

function which may be undertaken individually by [A] or [B] may also be undertaken by X'. (Quirk *et al*. 1985).

(2) A 3-day-old boy is brought to the emergency department because of a 6-hour history of [[rapid breathing] [and] [poor feeding]].

In general, linked units may comprise a wide range of grammatical categories and levels of syntactic projection.

In the present article, we adopt the terminology used by Quirk *et al*. (1985). Linked constituents are referred to as *conjoins*. Their linking forms a *coordinated constituent*. Overt linking of conjoins by coordinating conjunctions is referred to as *syndetic* coordination. Coordination in which the linking is not overtly marked, except by the occurrence of commas or semicolons in writing or tone unit boundaries in speech, is termed *asyndetic*.

Coordination usually links conjoins which are structurally and grammatically similar. As noted by Quirk *et al*. (1985), this can include some complex cases in which combined units such as indirect and direct objects (3), objects and direct complements (4), and objects and adverbials (5), are coordinated.

(3) We gave [[[William] [a book on stamps]] [and] [[Mary] [a book on painting]]].
(4) Jack painted [[[the kitchen] [white]] [and] [[the living room] [blue]]].
(5) You should serve [[[the coffee] [in a mug]] [and] [[the lemonade] [in a glass]]].

In the present article, these are considered examples of verb phrase (VP) coordination with ellipsis of the second head verb.

With regard to noun phrases (NPs), Quirk *et al*. describe *segregatory coordination* (2) and *combinatory coordination* (6). The two can be distinguished by considering the relationship of the coordinated constituent and its conjoins to its predicate. If the coordinated constituent can be replaced by each of its conjoins in the sentence and the meanings of the new sentences are consistent with that of the original, then the coordination is segregatory. If this replacement creates new sentences whose meanings are not consistent with that of the original (7), then the coordination in combinatory.

(6) The patient usually complains of [[pins] [and] [needles]] in the deltoid area.
(7) *The patient usually complains of [pins] in the deltoid area. The patient usually complains of [needles] in the deltoid area.

Due to the scarcity of combinatory coordination in the corpus described in Section 3.1 of this article, which provides evidence of the occurrence and use of coordination in this context, all NP coordination is considered segregatory in the research described here.

In writing, coordination is indicated by the use of conjunctions and punctuation. The approach taken in the current article focuses on *potential coordinators* which comprise the coordinating conjunctions *and*, *but*, and *or*, semicolons, commas, and adjacent comma–conjunction pairs. By definition, the coordinating conjunctions usually serve as coordinating links between conjoins. There is far more ambiguity in the use of commas, which may have either coordinating or subordinating functions.

Nunberg *et al*. (2002) describe the use of punctuation in English. They note that commas, semicolons, and colons normally mark constituent boundaries within sentences. In addition to coordinated units, commas serve to mark the boundaries of subordinated constituents such as post-modifiers (8), adverbial modifiers (1), and other subclausal constituents, which are less central to the main message being conveyed in the sentence (9). Nunberg *et al*. (2002) note that adjuncts, parentheticals, supplementary relative clauses, vocatives, and a range of others are all commonly bounded in this way by delimiting commas.

(8) [His father has schizophrenia[,] [paranoid type][,] treated with haloperidol and trihexyphenidyl].
(9) [Examination[,] [including cardiovascular examination[,]] shows no abnormalities].

In the context of our current work, the term *simple sentence* is used to denote declarative sentences containing no coordinated constituents. The aim of this research is to improve the performance of an IE system by means of a module that rewrites input sentences containing coordinated constituents as sequences of simple sentences. The module is also intended to recognize some types of subordinated constituent and exploit them in the IE process. One hypothesis tested in this article is that it is more effective for a system to exploit a small number of rules to extract pertinent facts from simple sentences than to exploit a larger number of rules in an effort to address the variation that results from coordination in natural language.

The detection of potential coordinators, their classification, and the identification of their conjoins is a prerequisite to realizing this aim. In the present article, we assume that coordination in a sentence can be detected by reference to potential coordinators. Given that the function of potential coordinators is often ambiguous, especially in the case of commas, it is necessary to recognize their use as subordinators and coordinators. Furthermore, as described in Section 3.1, coordination can hold between a variety of syntactic categories at various levels of syntactic projection. For this reason, once a system has identified a coordinator for the purpose of rewriting a complex sentence as a sequence of simple sentences, it is necessary for it to further identify the particular type of coordination signaled by the coordinator from the wide range of possibilities that exist.

Explicitly, the aim of the sentence rewriting module described in Section 3 is to convert sentences such as (10) into sequences of sentences such as (11).

(10) Examination shows [[jaundice][,] [hypothermia][,] [hypotonia][,] [[large [[anterior] [and] [posterior]]] fontanels][, and] [a hoarse cry]].

(11) Examination shows [a hoarse cry]. Examination shows [hypotonia]. Examination shows [large [anterior [fontanels]]]. Examination shows [large [posterior [fontanels]]]. Examination shows [jaundice]. Examination shows [hypothermia].

In this article, Section 2 motivates research into the rewriting of complex sentences as sequences of simple sentences for the purpose of an application in natural language processing (NLP), IE. The initial system is described and several performance issues noted. Section 3 begins with a description of the corpus of clinical vignettes that serves as the basis for the sentence simplification method presented in this article. An analysis of this corpus is described and findings regarding the use and the range of types of coordination and subordination that occur in it is presented. This section also presents a new machine learning classifier for potential coordinators in natural language sentences. It automatically labels instances as belonging to one of a wide variety of subordinating or coordinating classes derived from the analysis of the corpus. A range of baseline classifiers are also described. Finally, Section 3 describes an algorithm for rewriting complex sentences into sequences of simple sentences that exploits the classifiers. Section 4 presents related work on coordination, punctuation, its automatic treatment, and exploitation in NLP. Evaluation of the new approaches is presented in Section 5, which includes a comparison of IE systems exploiting the classifiers and sentence rewriting module described in Section 3. Section 6 presents plans for future work while Section 7 discusses the findings of the article and draws some conclusions. Throughout the article, unless stated otherwise, all linguistic examples are drawn from the corpus of clinical vignettes presented in Section 3.1. Relevant conjoins, coordinators, and subordinators are delimited using square brackets.

2 Motivation: IE from Clinical Vignettes

The research described in this article was undertaken in the context of a project on IE from vignettes that provide brief clinical descriptions of patients. The discourse structure of these vignettes consists of seven elements:

(1) Basic information;
(2) Chief complaint;
(3) History;
(4) Vital signs;
(5) Physical examination;
(6) Diagnostic study; and
(7) Laboratory study

Considering each in turn, *Basic information* describes the patient's gender, profession, ethnicity, and health status. *Chief complaint* presents the main concern that led the patient to seek therapeutic intervention. *History* is a narrative description of the patient's social, family, and medical history. *Vital signs* are a description of the patient's pulse and respiration rates, blood pressure, and temperature. *Physical examination* is a narrative description of clinical findings observed in the patient. *Diagnostic study* and *Laboratory study* present the results of several different types of clinical test carried out on the patient.

Each element in the discourse structure is represented by a template encoding related information. For example, the template for physical examinations holds information on each clinical finding or symptom (*finding*) observed in the examination, information on the technique used to elicit that finding (*technique*), the bodily location to which the technique was applied (*location*), the body system that the finding provides information on (*system*), and any qualifying information about the finding (*qualifier*). In this article, we focus on automatic extraction of information pertaining to physical examinations. The goal of the IE system is to identify the phrases used in the clinical vignette that denote findings and related concepts and add them to its database entry for the vignette.

In the research described in this article, the IE system depends on several NLP modules:

(1) Sentence tagger;
(2) Concept tagger; and
(3) Relation extractor.

Modules 1 and 2 are arranged in a pipeline, each one adding XML annotation to its input and passing this on to be exploited by the next module. Both were developed in house. The concept tagger uses gazetteers to tag references to clinical concepts mentioned in the vignette. In light of the specificity of the IE task undertaken in this research, the gazetteers were developed in-house on the basis of corpus analysis. Existing resources such as SNOMED and UMLS were considered, but their size and scope made them difficult to exploit in the current research. Hand-crafted finite-state transducers were used in conjunction with the gazetteers to group sequences of adjacent concepts together.

With regard to the third module in the IE pipeline, two relation extraction modules, BASIC and PATTERNS, were implemented for the purpose of comparison. Both of them exploit the annotation of sentences and clinical concepts obtained from the first two modules.

BASIC consists of a small number of simple rules. To summarize briefly, vignettes are processed by considering each sentence in turn. The first clinical finding or symptom mentioned in a sentence is taken as the basis for a new database entry. Similarly, the first tagged technique, system, and location within that sentence is considered to be related to the finding. Qualifiers (e.g. *bilateral* or *peripheral*) are extracted in the same way, except in sentences containing the word *no*. In these cases, the qualifier related to the finding is identified as *none*. Due to their scarcity in the corpus, this rule was not extended to additional negative markers such as *never* or *not*. When processing sentences generated by the simplification module described in Section 3.3, if the input sentence contains no tagged techniques, then BASIC attempts to extract this information from any adverbials identified in the sentence.

The PATTERNS relation extraction module takes every mention of a finding or symptom tagged in the input vignette as the basis for a new physical examination entry in the database. A set of hand-crafted rules is then applied to identify references to related concepts mentioned in the vignette. By way of illustration, references to the location to which a technique is applied in order to elicit a clinical finding are identified by selecting, on the basis of the first applicable rule, the tagged location occurring in any of the patterns:

1. ***technique*** {*at*|*over*} *the* _
2. ***technique*** *of the* _
3. _ ***system*** *is* ***finding***
4. ***finding*** . . . {*of*|*at*|*over*} *the* _
5. _ ***technique*** {is|are} ***finding***
6. ***finding*** *in the* ***qualifier*** _
7. _ *is* ***finding***
8. _ ***finding***

An underbar is used to indicate the position of the location in each pattern. Similar rule sets are used in the identification of the other concepts related to the finding. For brevity, they are not presented in this article.

The hand-crafted IE rules exploited by the PATTERNS module are implemented using regular expressions. They are applied in order, exploiting lexical and conceptually tagged elements. Quantitative evaluation of the BASIC and PATTERNS IE systems is presented in Section 5. In this section, we make some general observations on the outputs of the latter system.

It was noted that many errors were caused by its inability to accurately process coordinated constituents. Consider (12) and (13).

(12) Physical examination shows [[enlarged supraclavicular nodes that are stony hard] [, and] [a liver that is [[enlarged] [and] [irregular]]]].
(13) Examination of [the [[heart][,] [lungs] [, and] [abdomen]]] shows normal findings.

In (12), noun phrase and adjectival coordination means that there is a mismatch in the number of explicitly mentioned concepts: three findings, one technique, and two qualifiers. In (13), coordination of the head nouns causes a similar mismatch in the numbers of explicitly mentioned concepts: one finding, one technique, and three systems. The rules implemented in the initial IE system cannot detect the ellipsis of elements that occurs due to this coordination and are unable to reliably identify the relations holding between explicitly mentioned and elided concepts.

The patterns exploited by this initial IE system are too simple to accurately detect the relations that hold between the concepts tagged in these sentences. While the use of additional regular expressions would enable more accurate processing of them, they would be of limited use beyond the specific cases that they were designed to address. The variability of input sentences due to syntactic coordination is so great that it should be handled systematically rather than heuristically. Attempting to meet this challenge by the formulation of additional IE patterns would lead only to small improvements and would be a continual process. This line of reasoning motivates the development of the systematic approach to coordination presented in Section 3.

3 An Automatic Treatment of Coordination

This section presents a method to automatically rewrite sentences containing potential coordinators as sequences of simple sentences. It relies on a corpus in which potential coordinators have been annotated with information about their specific coordinating or subordinating function. The annotation is exploited by methods to classify previously unseen potential coordinators. Finally, a sentence simplification algorithm utilizing the classifiers is presented.

3.1 An annotated corpus

A corpus consisting of 138,641 words from 708 clinical vignettes was compiled in order to support development and evaluation of the IE system described in Section 2. The vignettes are written in academic US English and are highly consistent in their use of terminology, punctuation, and grammatical style.

Potential coordinators, including conjunctions, commas, and adjacent comma–conjunction pairs, were manually annotated in this corpus. In this article, seven types of potential coordinator are considered: *and*, *but*, *or*, *comma*, *comma-and*, *comma-but*, and *comma-or*. The decision to treat comma–conjunction pairs separately from commas or conjunctions alone was made on the basis that they usually introduce the final conjoin of coordinated constituents. It is thus likely that they share contexts distinct from those of the other potential coordinators.

The annotated corpus was divided into a training portion and a testing portion. The characteristics of the two are presented in Table 1.

In order to address our aim of implementing a module to automatically rewrite complex sentences for the purpose of subsequent NLP tasks, it is important to identify the different roles that may be played by potential coordinators. Instances of potential coordinators occurring in the corpus of vignettes were manually annotated with labels indicating their function. Where an instance occurs between two conjoins, its label conveys information about those conjoins.

The different classes of instance are divided into two sets, one for coordinators and another for subordinators. Tables 2 and 3 display the different classes of coordinator and subordinator annotated in the training corpus. The abbreviations used in the tables consist of a minimum of three components. The first indicates whether the class has a coordinating (C) or subordinating (S) function. The second component indicates the projection level of the constituents: morphemic (P), lexical (L), intermediate (I), maximal (M), or clausal/extended (C). The third element of each acronym is an abbreviation of the grammatical category of the constituents: nominal (N), verbal (V), adjectival (A), adverbial (Adv), prepositional (P), quantificational (Q), or unclear (X). A final numerical value is used to differentiate classes that cannot be distinguished on the basis of the criteria previously listed. To illustrate, CMV2-6 denote coordination of VPs in which the head of the rightmost VP has been elided and the conjoined VPs have distinct argument structures, as in sentences (3) to (5). The adoption of such specific classes is expected firstly to enable automatic classifiers to leverage very specific patterns of PoS tags, words, and semantic concept labels in their recognition and secondly, to enable each class to be associated with specific and accurate sentence simplification patterns.

Table 1 Characteristics of the annotated corpus

Characteristic	Training	Testing
#Items	422	286
#Words	107,900	30,741
#Sentences	12,451	3286
#Potential coordinators	4709	1491

The annotated corpus described here serves as the basis for development and evaluation of the classification modules for potential coordinators described in Section 3.2.

Table 2 Classes of coordinator occurring in the training and testing corpora

Category	Morphemic	Lexical	Intermediate	Phrasal	Clausal
Noun		CLN	CIN	CMN1, CMN2, CMN3	
Verb		CLV		CMV1, CMV2, CMV3, CMV4, CMV5, CMV6	CCV
Adjective	CPA	CLA		CMA1, CMA2	
Adverb				CMAdv	
Preposition		CLP		CMP	
Quantifier		CLQ			
Miscellaneous				CMM1, CMM2	

Table 3 Classes of subordinator occurring in the training and testing corpora

Category	Morphemic	Lexical	Intermediate	Phrasal	Clausal
Noun				SMN	
Verb					
Adjective				SMA	
Adverb				SMAdv1, SMAdv2	
Preposition					
Quantifier					
Miscellaneous				SMM1, SMM2	SCM

3.2 Automatic classification of potential coordinators

Several methods were implemented for the automatic classification of potential coordinators. These classifiers are described in Sections 3.2.1–3.2.4. Their evaluation is presented in Section 5.

3.2.1 *Memory-based learning classifier*

Instances of potential coordinators in the training corpus were processed in order to represent them as vectors of feature values encoding their linguistic properties and context of use. The initial representations exploited sixty-five features. A classifier of potential coordinators was derived from this training data using the TiMBL memory-based learner (MBL) (Daelemans *et al.*, 2010). Feature selection and algorithm optimization were performed using a simple hill-climbing procedure.

The initial set of features encodes different kinds of information about each potential coordinator. They can be grouped as follows:

(1) orthographic form of the potential coordinator;
(2) information on the position of the instance within the document;
(3) information about items that both precede and follow the potential coordinator within the same sentence. This includes:
    (a) words and their parts of speech;
    (b) clinical concepts;
    (c) the number of determiners;
    (d) the distance in words to the next following determiner if a determiner also precedes the instance; and
    (e) the parts of speech that immediately precede and follow *other* potential coordinators that both precede and follow the instance.
(4) Boolean features asserting various conditions that hold over items occurring in the same sentence as the potential coordinator:
    (a) words with matching parts-of-speech *p* precede and follow the instance. Here, *p* comprises verbs of the past, past participle, and singular present tenses, determiners, cardinal numbers, adjectives, pronouns, and nouns;
    (b) an adverb precedes the instance;
    (c) the instance is both preceded and followed by a word with part-of-speech *q:*
        (i) where *q* includes adjectives and past participle verbs; and
        (ii) where *q* includes potentially mismatched singular or plural common nouns or proper nouns.
    (d) The instance is *immediately* preceded and followed by a word with part-of-speech *q*, where *q* is:
        (i) determiner; and
        (ii) cardinal number.
    (e) The words *no*, *not*, or *either* precede the instance in the sentence.
    (f) An adverb or preposition precedes the instance; and
    (g) Textual material that includes a word with part-of-speech *p* followed by a word with part of speech *q* both precedes and follows the instance where:
        (i) *p* is an adjective and *q* is a preposition;
        (ii) *p* is nominal and *q* is an adverb;
        (iii) *p* is a cardinal number and *q* is a preposition; and
        (iv) *p* is nominal and *q* is a preposition.
(5) A domain-specific ternary feature indicating whether the potential coordinator is either preceded or followed by the word *history* in the sentence, or both preceded and followed by that word; and
(6) Features that combine the values of another pair of features into a single feature:
    (a) immediately preceding and following part-of-speech tags (built from features in 3.a); and
    (b) closest preceding and following conceptual tags (built from features in 3.b).

Table 4 displays the groups of feature selected for the classification of each type of potential

Table 4 Features selected for optimal classification of different potential coordinators

Potential coordinator	Feature group	Proportion of features selected
and	**3.a**, 3.b, 3.c, **3.e**, 4.a, 4.c.i, 4.d.ii, 4.f, 4.g.ii, 4.g.iii, 4.g.iv, 5, 6.a, 6.b	0.4923
but	2, **3.a**, 3.b, 3.c, **3.e**, 4.a, 4.b, 4.c.i, 4.d, 4.f, 6.a, 6.b	0.4154
or	**3.a**, **3.e**, 4.c.ii, 6.a	0.2308
comma	**3.a**, 3.b, 3.c, **3.e**, 4.a, 4.c.i, 4.f, 5, 6.a, 6.b	0.4154
comma-and	**3.a**, 3.b, 3.c, 3.d, **3.e**, 4.a, 4.c.i, 4.c.ii, 4.f, 4.g.iii, 5, 6.a	0.5231
comma-but	**3.a**, **3.e**, 6.a	0.0461
comma-or	3.d, **3.e**, 6.a	0.0461

coordinator. The third column shows the proportion of features from the initially proposed set that are selected for optimal classification accuracy. The most globally important feature groups appear in bold font in Table 4.

The optimization revealed that for all potential coordinators, TiMBL worked best when using the TRIBL2 algorithm. Table 5 presents the optimal settings for other parameters with respect to each potential coordinator. Instances occurring in input data are classified using TiMBL with these optimal parameter settings. Section 5 presents an evaluation of the optimized classifiers described here.

3.2.2 *Stanford parser classifier*

This classifier (STANFORD) exploits the Stanford Lexicalized Parser v1.6.3 (Klein and Manning, 2003) for the purpose of classifying potential coordinators. The constituent structure returned by the parser can be used to derive a classification for potential coordinators for most of the classes presented in Table 2. A set of simple conversion rules exploiting regular expressions was employed for this purpose. The classification of subordinators is slightly more difficult, and is based on the recognition of patterns in the upper nodes of the tree output by the parser.

3.2.3 *Hybrid classifier*

Evaluation of the two classifiers over the testing data showed that the MBL and STANFORD classifiers have somewhat orthogonal performance. This motivated the development of a hybrid classifier (HYBRID) that uses the MBL classifier when processing conjunctions and commas and uses the STANFORD classifier when processing adjacent comma–conjunction pairs.

Table 5 Optimal algorithm parameter settings when classifying potential coordinators

Parameter setting		Classifiers
Feature weighting	Gain ratio	*and*, *but*, *or*, *comma*, *comma-but*
	Shared variance	*comma-or*
	No weighting	*comma-and*
Class voting weight	Inverse distance	*and*, *but*, *or*, *comma*, *comma-or*
	Normal majority voting	*comma-and*, *comma-but*
Distance metric	Modified value difference	*and*, *but*, *or*, *comma*, *comma-and*
	Jeffrey divergence	*comma-or*
	Overlap	*comma-but*
Neighbours	3	*comma-and*, *comma-but*
	4	*but*, *or*, *comma*
	5	*and*
	21	*comma-or*

3.2.4 *Majority class baseline classifier*

This baseline classifier (MAJORITY) is based on observation of the frequency with which different classes of coordination and subordination occur in the training corpus. It classifies every instance with the most frequently observed class for potential coordinators of that type. Every instance of *and* and *comma-or* are classified as CMN1, every instance of *but* and *or* is classified as CMV1, every *comma* is classified as SMAdv1, and every instance of

```
Input: Sentence containing coordinated constituents, s_0
Output: Array of simple sentences, A; Adverbial modifier, adv
A ← ∅;
adv ← empty string;
S ← {s_0};
while S ≠ ∅ do
    s_i ← pop(S);
    if s_i contains a coordinator/subordinator of a type listed in Table 6 then
        (adv, §_i) ← simplify(s_i);
        ∫_i ← dereference(§_i);
        S ← S ∪ {∫_i};
    else
        A ← A ∪ {s_i}
    end
end
```

Fig. 1 The sentence simplification algorithm

comma-and and *comma-but* is classified as CCV under this method.

3.3 Sentence rewriting

The syntactic simplification method exploits all the annotations added to input sentences by the previously described modules. A part of speech tagger (Brill, 1994) is also exploited. The method is based on a recursive algorithm operating over an array of sentences (Fig. 1). In its initial state, this array comprises a single sentence containing one or more instances belonging to any of the ten classes displayed in Table 6.

The sentence simplification algorithm is presented in Fig. 1. The function *simplify* consists of an ordered set of quick-fire rules, designed to process different classes of coordination and subordination indicated by different types of instance. Each rule identifies a coordinator/subordinator, t_i, and generates a pair of sentences, $\int_i^1$ and $\int_i^2$ The former is derived from textual material preceding t_i in the input sentence, while the latter is derived from material following it. The function returns any identified adverbials, *adv*, and a reference, $\S_i$, to the pair of generated sentences. The rules were developed manually by reference to the test corpus and a key file containing information on the class of each potential coordinator.

Table 6 Classes of coordinator/subordinator triggering simplification rules

Coordinator/ Subordinator	Classes
and	CCV, CMN1, CIN, CLA, CMA1, CMV1
but	CMN1, CMA1, CMV1
or	CMN1, CIN, CLN, CMV1
comma	SMAdv1, CCV, SMM1, SMM2, CMN1, CMA1, CLA
comma-and	CMV1, CMN1, CLN, CCV
comma-but	CMA1, CCV
comma-or	CMN1

To illustrate with two examples of rules:

- SMAdv1 triggers a rule that recognizes preceding material as an adverbial modifier of the input sentence, *adv*. $\int_i^1$ is an empty string and $\int_i^2$ is the part of the sentence that follows t_i. A binary array consisting of $\int_i^1$ and $\int_i^2$ is built.
- When instances of class CMN1 occur in a context such as A B/vbz C t_i D in s_i, a rule is triggered which constructs an array consisting of the strings $\int_i^1$: A B/vbz C and $\int_i^2$: A B/vbz D. If s_i is (14), this rule derives (15) as $\int_i^1$ and (16) as $\int_i^2$. The upper case letters A–D are regular expressions matching text intervening between the

Table 7 Characteristics of rewrite rules by class

Class	Coordinator/ subordinator	#Rewriting rules	Order of precedence
CCV	*and*	1	3
	comma	1	4
	comma-and	1	2
	comma-but	1	2
CMN1	*and*	26	8
	but	1	9
	or	9	10
	comma	7	12
	comma-and	7	11
	comma-or	1	11
CIN	*and*	2	17
	or	1	18
CLA	*and*	2	21
	comma	2	22
CMA1	*and*	1	14
	but	1	14
	comma	2	16
	comma-but	1	15
CMV1	*and*	2	6
	but	2	6
	or	2	6
	comma-and	1	7
CLN	*or*	1	19
	comma-and	1	20
SMAdv1	*comma*	1	1
SMM1	*comma*	2	5
SMM2	*comma*	2	5

strings specified within the pattern. vbz is a part of speech tag denoting present tense verbs. The fact that the NP conjoins follow the verb implies that they form a coordinated object. This assumption motivates the form of the rule.

(14) She has diabetic retinopathy [but] no evidence of renal disease.
(15) She has diabetic retinopathy.
(16) She has no evidence of renal disease.

The output, *A* is a set of sentences containing no instances of the types listed in Table 6. The IE function is applied to this set of sentences and any adverbial information derived during the rewriting process.

Table 7 displays characteristics of the rewrite rules used by the simplification algorithm. Column 3 of the table displays the possible number of rewrite rules that may be applied by the algorithm on encountering different types of coordinator/subordinator. This information serves as an indirect indicator of the challenge posed by each rewriting task. It can be noted that many more rules are needed to cater for the various functions and contexts of NPs in the clinical vignettes than other types of constituent. The rules include heuristics that exploit PoS tagging and preposition and verb recognition to identify the syntactic function of coordinated NPs. Column 4 of Table 7 shows the relative order in which the rules are tested against an input sentence. The success of the rewriting algorithm depends on both the patterns exploited by the rules and their order of application. In general, the rules are intended to process the coordination of larger and more syntactically dominant conjoins first.

The simplified sentences derived by the interaction of this module with each of the classifiers of potential coordinators described in Section 3.2 are then processed by the BASIC IE system described in Section 2. The templates produced by this IE system and the PATTERNS IE system, described in the same section, are evaluated in Section 5.

4 Related Work

In conducting the research presented in this article, a review of previous related work was undertaken. This includes research on the disambiguation of coordinated structures and the role of punctuation and research describing the exploitation of information about coordination in syntactic parsing, IE, and other NLP applications.

Addressing the challenge of disambiguating and processing coordination, Agarwal and Boggess (1992) present a system to identify the boundaries of conjoins linked by coordinating conjunctions. Their rule-based algorithm exploits concept tagging, part of speech tagging, and the use of a ‘semi-parser’ to identify constituents such as NPs, VPs, and PPs. It performs with an accuracy of 81.6%, but is noted to be unable to identify clausal conjoins and does not recognize coordination indicated by commas.

Many of the approaches presented in the literature recognize that there is likely to be syntactic and semantic similarity between conjoins involved in coordination and exploit this in order to disambiguate coordinate structures (Kurohashi and Nagao, 1992; Resnik, 1999; Goldberg, 1999; Chantree *et al*., 2005).

Buyko and Hahn (2008) sought to learn the extent of the contribution made by the recognition of semantic similarity between conjoins to the processing of coordination. They found that a system based on conditional random fields exploiting semantic features was outperformed by one based on output from a syntactic parser. In the present article (Section 3.2.1), features encoding semantic information were selected for exploitation by several classifiers, though the statistical significance of their contribution has not been assessed. Shimbo and Hara (2007) describe an approach to the disambiguation of coordinate conjunctions based on methods from sentence alignment. Their system was found to outperform state-of-the-art parsers when processing the GENIA Treebank beta corpus.

Kawahara and Kurohashi (2007) present methods to disambiguate coordination in Japanese. Exploiting verb case frames automatically derived from the web, the method applies lexical preferences and co-occurrence statistics between potential conjoins to resolve coordination ambiguities. An updated approach exploiting functional dependency information was described in Kawahara and Kurohashi (2008).

Methods exploiting information about neighbouring syntactic constituents have been used to disambiguate the role of commas in natural language. This work was described in Bayraktar *et al.* (1998) and Srikumar *et al.* (2008).

With regard to IE, Rindflesch *et al.* (2000) used an automatic treatment of coordination to improve IE of facts about macromolecular binding. In a contrasting approach, Klebanov *et al.* (2004) present a method to improve performance in IE without processing coordination. Their approach relies on the identification of 'easy-access sentences' (EAS) that contain a single finite verb in a 'semantically non-problematic environment' and a large number of named entities (concepts). In this approach, IE rules are applied only to EASs. The accuracy with which EASs are identified is reported in this work, but unfortunately changes in accuracy elicited in their IE system as a result of applying the method are not presented.

Many authors demonstrate that the use of methods to improve the resolution of coordination ambiguities improves overall performance in syntactic parsing for various languages (Ratnaparkhi *et al*., 1994; Rus *et al*., 2002; Kim and Lee, 2003; Charniak and Johnson, 2005; Nakov and Hearst, 2005; Hogan, 2007; Kübler *et al*., 2009). In addition to this, various papers report on the exploitation of information about coordination for other tasks in NLP and in industrial contexts. Rindflesch (1995) incorporated a method for dealing with coordination to improve the mapping of NPs identified in input documents to concepts in the medical UMLS database. Cederberg and Widdows (2003) present a method exploiting information about noun coordination to improve automatic hyponymy extraction. The method is based on well-established lexicosyntactic patterns modified to allow recognition and exploitation of coordinated structures. The hyponymy relations identified are then filtered using latent semantic analysis. In their preliminary work, Tjong and Berry (2008) seek to improve the clarity of industrial requirements specifications by minimizing the ambiguous use of coordination. Their paper describes a range of semantic relations implied by the use of coordination and urges the adoption of rules similar to a controlled language specifying a writing policy for coordinated structures.

Despite the amount of work addressing the issue of coordination in natural language, the contribution brought by these approaches to practical NLP applications has been little reported. Overall, the work surveyed in this section was useful in guiding development of the features presented in Section 3.2.1. Authors have drawn differing conclusions as to the suitability of different types of information in resolving coordination ambiguities. This observation motivated the approach adopted in the present article, in which an initial feature set is developed and a feature selection method is applied in order to derive the optimal subset to be exploited by the classifier.

5 Evaluation

This section presents an evaluation of the modules described in Section 3. In all cases, where comparisons are made between different systems in terms of accuracy or *F*-score, significance was computed using approximate randomization (Chinchor, 1992). The significance threshold, $\alpha = 0.05$.

The production of annotated data for the task of sentence simplification is costly and complex. For this reason, different settings of the sentence simplification module will be evaluated extrinsically (Sparck-Jones and Galliers, 1996) via the performance of the IE system that exploits them.

Unfortunately, due to the nature of the sources from which information is to be extracted in this work, no direct comparison can be made with the systems presented in previous research. Recognizing this problem, the new modules presented in this article are compared with one based on the publicly accessible Stanford parser.

5.1 Evaluation of the classification of potential coordinators

Table 8 presents the accuracy scores of the different classifiers obtained using ten-fold cross-validation over the training corpus. Given their superiority over the MAJORITY classifier, the main focus of this section will be in comparing the accuracy of the MBL and STANFORD classifiers. It can be observed that the MBL classifier classifies all potential coordinators except *comma-but* with greater accuracy than the STANFORD classifier. However, the only potential coordinators for which there was a statistically significant difference in performance were the two most common, *comma* and *and*.

A detailed class-by-class examination of the performance of the MBL classifier reveals that for all potential coordinators, the most common type of error concerns the projection level of nominal constituents. This finding was also derived from an analysis of inter-annotator agreement over a sample of the training data. It can be noted that errors made by the STANFORD classifier are similar in kind, though there is more evidence of the erroneous assignment of grammatical category as well as projection level to coordinated constituents.

Table 8 Classification accuracy obtained via ten-fold cross-validation over the training set

Potential coordinator	#Instances	MAJORITY	STANFORD	MBL
and	1544	0.3543	0.5971	0.7506
but	100	0.7100	0.8000	0.8700
or	80	0.2625	0.4750	0.6000
comma	1931	0.2952	0.7369	0.8716
comma-and	965	0.6953	0.8788	0.8891
comma-but	75	0.9600	0.9867	0.9733
comma-or	14	0.5714	0.5714	0.7143
ALL	4709	0.4158	0.7205	0.8320

Overall, of the eighty-one combinations of classes and types, the *F*-score obtained by STANFORD is superior to MBL in thirteen. The most frequent of these thirteen is CCV signalled by *comma-and*, which accounts for 14.25% of all instances annotated in the training data. However the margin of difference is slight (*F* 0.9880 versus 0.9719), and the contribution of this improvement to the IE system is not envisaged to be great. A similar description can be made with regard to the greater *F*-score obtained by STANFORD with regard to the CCV class signalled by *comma-but*, which accounts for 1.53% of the training data. There are classes for which STANFORD obtains a significantly higher *F*-score than MBL, but each of these accounts for less than 1% of the total training set.

One reason for the relatively poor performance of STANFORD is that the labels returned by the Stanford parser are not as specific as those used in the manual annotation of the training data exploited by MBL. To illustrate, the label VP used by the Stanford parser subsumes two classes (CMV1 and CMV2) and NP subsumes three (CMN1, CIN, and CMV3). The STANFORD classifier is therefore unable to differentiate between these classes.

Finally, it has been noted that both MBL and STANFORD classifiers fail to identify instances of class CIN (17). The most common type of error involving CIN is misclassification as CMN1. The simplification rules applied to these classes are similar in many ways, relying on identification of nominal and verbal heads in the sentence. It is therefore expected that such errors will not be too detrimental.

Table 9 Classification accuracy obtained over the test set

Potential coordinator	#Instances	MAJORITY	STANFORD	MBL
and	137	0.3650	0.4453	0.6642
but	13	0.3077	0.5385	0.7692
or	12	0.0833	0.4167	0.4167
comma	91	0.1209	0.5604	0.7363
comma-and	49	0.3673	0.8163	0.7347
comma-but	6	0.6667	0.8333	0.6667
comma-or	2	0.5000	0.5000	0.0000
ALL	310	0.2871	0.5484	0.6871

(17) The sclerae and the skin of the [[head] [and] [upper trunk]] are yellow.

Table 9 presents the classification accuracy of different classifiers when processing a subsample of the test corpus which consists only of sentences that mention clinical findings. Over the classes of coordination and subordination occurring in descriptions of physical examinations, STANFORD classifies potential coordinators consisting of adjacent *comma–conjunction* pairs more accurately than MBL does. The difference in classification accuracy between the STANFORD and MBL classifiers is statistically significant with regard to the potential coordinators *comma-but*, *and*, and *comma*. In classifying the latter two types, MBL is superior whereas in classifying the first, STANFORD is superior.

5.2 Evaluation of IE exploiting classification of potential coordinators and sentence simplification

Testing data for the IE task was derived from the stems of 70 clinical vignettes. The set contains 206 clinical findings and related concepts. The IE systems, PATTERNS and BASIC, described in Section 2 were used to process this data set. Several variants of the BASIC system were employed, each exploiting one of the different methods for classification of potential coordinators described in Section 3.2 and the sentence simplification algorithm presented in Section 3.3. A variant of the PATTERNS relation extraction module was also implemented that extracts just a single tagged finding from an input sentence as opposed to all tagged findings.

The metrics used in evaluation of the IE systems are based on accuracy. For findings, when the IE system identifies a finding within a particular sentence of a particular vignette, and the same finding has been marked within the same sentence of the same vignette in the key, this is considered a true positive. The accuracy score for findings is the ratio of the number of true positives to the total number of findings marked in the key. It is computed in a similar way for the concepts related to findings. Due to the strong semantic typing involved in the IE task and the limited number of candidates for selection with regard to a particular finding, accuracy was considered a more suitable metric than *F*-measure. The evaluation described here is based on exact string matching. Systems are not rewarded for obtaining partial matches.

Tables 10 and 11 display the accuracy of different IE systems in identifying clinical findings and related concepts in descriptions of physical examinations. Table 10 shows the performance of IE systems implemented only to identify the first tagged finding and concepts related to that finding in input sentences. Table 11 provides evaluation results for IE systems that identify all tagged findings and concepts related to those findings in input sentences.

In both tables, the columns IGNORE contain accuracy scores for systems that exploit the sentence rewriting module described in Section 3.3 but do not exploit any classification of potential coordinators. As a result, for these systems, the sentence rewriting rules are never activated. The columns MBL, STANFORD, HYBRID, and MAJORITY present accuracy scores for IE systems that work in the same way, but which exploit the classification modules for potential coordinators described in Sections 3.2.1–3.2.4, respectively. The columns PATTERNS present the accuracy scores obtained by the IE system described in Section 2.

The results are broadly in line with expectation. A comparison of the overall accuracy of the PATTERNS systems with the others supports the hypothesis that it is more effective to apply a small set of IE rules over simplified input sentences than to employ a larger set of complex IE rules in an effort to handle the variation exhibited by sentences containing coordinated constituents. For IE systems

Table 10 Accuracy of IE systems exploiting different classifiers of potential coordinators (assuming one finding per sentence)

Template slot	One finding per sentence						
	IGNORE	MAJORITY	PATTERNS	STANFORD	MBL	HYBRID	KEY
finding	0.5845	0.7584	0.5556	0.7971	0.7971	0.8019	0.8696
technique	0.7729	0.7681	0.7778	0.7778	0.7778	0.7874	0.8019
system	0.7536	0.8357	0.7391	0.8309	0.8406	0.8454	0.8744
qualifier	0.6812	0.8164	0.5797	0.7971	0.8261	0.8213	0.8261
location	0.8696	0.8889	0.8985	0.8985	0.9034	0.9082	0.9324
ALL	0.7324	0.8135	0.7101	0.8203	0.8290	0.8328	0.8609

Table 11 Accuracy of IE systems exploiting different classifiers of potential coordinators (assuming multiple findings per sentence)

Template slot	Multiple findings per sentence						
	IGNORE	MAJORITY	PATTERNS	STANFORD	MBL	HYBRID	KEY
finding	0.9420	0.8068	0.8454	0.8744	0.8551	0.8696	0.9275
technique	0.7729	0.7681	0.8116	0.7778	0.7778	0.7874	0.8019
system	0.7536	0.8357	0.8406	0.8309	0.8406	0.8454	0.8744
qualifier	0.6811	0.8164	0.6135	0.7971	0.8261	0.8213	0.8261
location	0.8696	0.8889	0.9130	0.8985	0.9034	0.9082	0.9324
ALL	0.8039	0.8232	0.8048	0.8357	0.8406	0.8464	0.8725

identifying single findings in input sentences, the fact that IGNORE is more accurate than PATTERNS was unexpected. However, it was found that the PATTERNS approach works significantly better if multiple findings are extracted from input sentences.

In Table 11, the IGNORE system is the most effective one at identifying findings mentioned in clinical vignettes. This suggests that even after syntactic simplification, some test sentences still mention multiple findings. Another possibility is that some coordinated constituents have been erroneously identified as findings in the key file.

A significance matrix was computed to plot a pair-wise comparison of all systems presented in this article. With $\alpha = 0.05$, the systems can be ranked as follows:

(1) KEY (multiple findings identified per sentence);
(2) KEY (one finding identified per sentence) and HYBRID (multiple findings per sentence);
(3) HYBRID and STANFORD (one finding per sentence) and MBL and STANFORD (multiple findings per sentence);
(4) IGNORE and PATTERNS (multiple findings per sentence) and MBL (one finding per sentence);
(5) MAJORITY (in both contexts);
(6) IGNORE (one finding per sentence); and
(7) PATTERNS (one finding per sentence).

This ranking is based on a comparison of the number of systems that a given system significantly outperforms with the number that significantly outperform it.

Although not statistically significant in this setting, the difference in accuracy between the MAJORITY and IGNORE systems in Table 10 shows that performance in IE can be improved even when the classification of potential coordinators is quite inaccurate.

5.3 Error analysis

The output of different modules within the IE system was examined in order to investigate the causes and impact of the errors they make. In this section, categories of errors are categorized as concerning conceptual tagging, the classification of potential coordinators, the simplification of sentences, and IE.

A number of errors arose as a result of the conceptual tagging process. In particular, there are clinical findings involving clinical procedures that were not included in our existing gazetteers and could not be recognized (e.g. *requires intubation/ mechanical ventilation*). Another omission of this type involves general vocabulary such as the word *moves* in the finding *he moves all extremities to painful stimuli*. Finally, several findings are numerical and their recognition depends on processing context, as in the example *Deep tendon reflexes are 1+*.

One particular weakness of the conceptual tagger is its inability to resolve ambiguities between adjectives that belong to different concept types according to the context of use. To illustrate, the modifier *stony-hard* functions as a qualifier or finding whereas *palpable* functions as a qualifier or technique, depending on the context of use. One additional challenge in the processing of qualifiers is the decision of whether to tag them as separate elements or to merge them with adjacent concepts.

With regard to the classification of potential coordinators, learning curves were plotted to show the correlation between training set size and classification accuracy for each type of potential coordinator. Examination of the learning curves suggests that a minimum of 200 instances are required in order to obtain a representative sample of the use of each potential coordinator. The training corpus used in this study contains far fewer instances than this of the potential coordinators *but*, *or*, *comma-but*, and *comma-or*. It is suggested that the training sets for these items should be increased considerably before the accuracy scores of the different classifiers can be regarded as definitive.

In the IE task, several errors were caused by a misclassification of sentences conveying information about the medical history of the patient and those concerning the physical examination. One example of this type is (18). Such errors arise because the IE system exploits information on the occurrence of particular verbs in the present tense when classifying sentences in the vignette as ones which provide information on physical examinations. The verb used in the first clause of sentence (18) is also commonly used in descriptions of physical examinations.

(18) [[Needle biopsy shows papillary carcinoma][, and] [he undergoes total thyroidectomy]].

There are several instances of errors in the IE key file in which phrases denoting findings and qualifiers contain potential coordinators. These cases may have some impact on the accuracy scores obtained by the IGNORE system evaluated in this article. However, they are infrequent enough that their influence on the evaluation results reported in Section 5.2 is not expected to be significant. No instances of combinatory coordination were noted in the test data.

6 Plans for Future Work

The linguistic studies discussed in Section 1 and the error analysis presented in Section 5.3 motivate five directions in which development of the sentence simplification module presented in this article may proceed.

One non-trivial improvement that could be made to the sentence simplification module would be to classify coordination as having either a segregatory or a combinatory interpretation. No assessment has been made of the significance of this issue in the context of the current IE task, but one possible approach to this challenge would be a method exploiting very large unannotated corpora. Quirk *et al.* (1985) note that one way for linguists to distinguish between segregatory and combinatory coordination is to check the acceptability of sentences created by inserting the word *both* before the first conjoin. This operation would produce sentence (19) from sentence (6).

> (19) *The patient usually complains of both [[pins] [and] [needles]] in the deltoid area.

It may be possible, by examining the frequencies of such constructed sentences in very large corpora, to recognize combinatory coordination in input sentences. Empirical approaches comparing the frequency of occurrence of constructions in which the order of the conjoins is reversed may also be examined.

Nunberg *et al*. (2002) present a description of the role and use of other punctuation symbols besides the comma such as indicators of parenthesis, single and double dashes, single and double quotation marks, related punctuation indicators, and the pragmatic implications that arise from the interaction of various punctuation marks. The modules described in the current article do not address these phenomena. For the current IE task, this is not problematic, but it is envisaged that IE from sources such as medical journals, text books, and patient notes may benefit from future work on the simplification of sentences employing this wider range of punctuation symbols.

The MBL classifier of potential coordinators was optimized using a naïve hill-climbing procedure in which feature selection and algorithm optimization are treated independently. Methods for joint optimization of the two have been undertaken in previous work (Daelemans *et al*., 2003). Such approaches are more computationally expensive, often exploiting clusters of processors employing genetic algorithms. It has been shown that joint optimization leads to the derivation of significantly more accurate classifiers by undertaking a more thorough exploration of the possibility space defined by different parameter settings. It will be interesting to apply such approaches in future work in order to derive more effective classifiers of potential coordinators.

For the scenario described in Section 2, the recognition and use of specific verbs in the rules used by the IE system is not important. However, this is not true of IE in alternate scenarios in which pertinent facts are identified by reference to the verbs linking different concepts. In light of this, it will be beneficial to apply a methodology to ensure subject–verb concord in the sentences generated by the module described in Section 3.3. This will ensure that a sentence such as (20) will be rewritten as a sequence such as (21) rather than (22). This improvement can be made using relatively simple morpho-syntactic rules.

> (20) [[Pelvic examination] [and] [urinalysis]] show no abnormalities.
> (21) [Pelvic examination] shows no abnormalities. [Urinalysis] shows no abnormalities.
> (22) *[Pelvic examination] show no abnormalities. [Urinalysis] show no abnormalities.

In addition to the expansion of the annotated corpus motivated by observations made in Section 5.3 and with improvement in subject–verb concord, it will be interesting to assess the contribution of the simplification process in other NLP applications such as question answering, pronoun resolution, multiple-choice question generation, and IE in different scenarios.

Finally, analysis of documents from alternate domains shows evidence of classes of subordination absent from the corpus described in Section 3.1. To be effective when applied to different domains, the annotation scheme for subordinators should be revised to include classes of comma signalling the left and right boundaries of different types of subordinated constituent.

7 Conclusion

Three main conclusions were drawn from the research described in this article. The first is that the automatic simplification of syntactic complexity can induce significant improvements in subsequent NLP tasks. A variety of approaches were implemented and evaluated by reference to the accuracy of an IE system exploiting them. Of the fully automatic modules tested, the best performing one was a hybrid system combining a memory-based learning classifier with a classifier derived from a syntactic parser. When exploiting classifiers based only on a syntactic parser or a memory-based learning method, sentence simplification still significantly improved the accuracy of the IE system.

The second conclusion to be drawn also follows from the comparative evaluation of variant IE systems. It was found that approaches that bypass a systematic treatment of coordination and handle coordination and subordination by means of more sophisticated IE rules perform relatively poorly.

The third conclusion to be drawn from this article follows from error analysis. It is expected that the syntactic simplification method described here will be improved by pursuing various lines of research. These include increasing the amount of annotated data available for development of some of the classifiers of potential coordinators, introducing procedures to disambiguate combinatory and segregatory coordination, developing a module to recognize the functions of a wider range of punctuation symbols in the simplification model, and introducing methods to ensure subject–verb concord in the sentences generated by the modules.

References

Agarwal, R. and Boggess, L. (1992). A Simple but Useful Approach to Conjunct Identification. *Proceedings of the 30th annual meeting for Computational Linguistics*. Newark, Delaware, June 1992.

Bayraktar, M., Say, B., and Akman, V. (1998). An analysis of English punctuation: The special case of comma. *International Journal of Corpus Linguistics*, **3**(1): 33–57.

Brill, E. (1994). Some Advances in Transformation-Based Part of Speech Tagging. *Proceedings of the Twelfth National Conference on Artificial Intelligence*. Seattle, Washington, July–August 1994.

Buyko, E. and Hahn, U. (2008). Are morpho-syntactic features more predictive for the resolution of noun phrase coordination ambiguity than Lexico-semantic similarity scores? *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, England, August 2008.

Cederberg, S. and Widdows, D. (2003). Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. Edmonton, Canada, May 2003.

Chantree, F., Kilgarriff, A., De Roeck, A., and Willis, A. (2005). Using a distributional Thesaurus to resolve coordination ambiguities. Technical Report 2005/02, The Open University.

Charniak, E. and Johnson, M. (2005). Coarse-to-Fine n-Best parsing and MaxEnt discriminative reranking. *Proceedings of the 43rd Annual Meeting of the ACL*. Ann Arbor, Michigan, June 2005.

Chinchor, N. (1992). The statistical significance of the MUC-4 results. *Proceedings of the Fourth Message Understanding Conference*. McLean, Virginia, June 1992.

Daelemans, W., Hoste, V., De Meulder, F., and Naudts, B. (2003). Combined optimization of feature selection and algorithm parameters in machine learning of language. *Proceedings of the 14th European Conference on Machine Learning (ECML-2003)*. Cavtat-Dubrovnik, Croatia, September 2003.

Daelemans, W., Zavrel, J., Van der Sloot, K., and Van den Bosch, A. (2010). TiMBL: Tilburg Memory Based Learner, version 6.3, Reference Guide. ILK Technical Report – ILK 10-01, ILK Research Group.

Goldberg, M. (1999). An Unsupervised Model for Statistically Determining Coordinate Phrase Attachment. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*. College Park, Maryland, June 1999.

Hogan, D. (2007). Coordinate noun phrase disambiguation in a generative parsing model. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Prague, Czech Republic, June 2007.

Kawahara, D. and Kurohashi, S. (2007). Probabilistic Coordination Disambiguation in a Fully-Lexicalised Japanese Parser. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech Republic, June 2007.

Kawahara, D. and Kurohashi, S. (2008). Coordination Disambiguation Without any Similarities. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, England, August 2008.

Kim, M.-Y. and Lee, J.-H. (2003). S-Clause Segmentation for Efficient Syntactic Analysis Using Decision Trees. *Proceedings of the Australasian Language Technology Workshop*. Melbourne, Australia, December 2003.

Klebanov, B. B., Knight, K., and Marcu, D. (2004). Text Simplification for Information-Seeking Applications.

In Meersman, R. and Tari, Z. (eds), *On the Move to Meaningful Internet Systems 2004*. Berlin: Springer-Verlag, pp. 735–47.

Klein, D. and Manning, C. D. (2003). Fast exact inference with a factored model for natural language parsing. *Advances in Neural Information Processing Systems (NIPS-15)*. Vancouver, British Columbia, December 2002.

Kübler, S., Hinrichs, E., Maier, W., and Klett, E. (2009). Parsing Coordinations. *Proceedings of the 12th Conference of the European Chapter of the ACL*. Athens, Greece, March 2009.

Kurohashi, S. and Nagao, M. (1992). Dynamic Programming Method for Analysing Conjunctive Structures in Japanese. *Proceedings of COLING-92*. Nantes, France, August 1992.

Nakov, P. and Hearst, M. (2005). Using the Web as an Implicit Training Set: Application to Structural Ambiguity Resolution. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*. Vancouver, British, Columbia, October 2005.

Nunberg, G., Briscoe, T., and Huddleston, R. (2002). Punctuation. In Huddleston, R. and Pullum, G. K. (eds), *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press, pp. 1724–64.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Ratnaparkhi, A., Roukos, S., and Ward, R. T. (1994). A Maximum Entropy Model for Parsing. *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. Yokohama, Japan, September 1994.

Resnik, P. (1999). Semantic similarity in a Taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *The Journal of Artificial Intelligence Research*, **11**: 95–130.

Rindflesch, T. C. (1995). Integrating Natural Language Processing and Biomedical Domain Knowledge for Increased Information Retrieval Effectiveness. *Proceedings of the Fifth Annual Dual-use Technologies and Applications Conference*. Utica/Rome, NY, May 1995.

Rindflesch, T. C., Rajan, J. V., and Hunter, L. (2000). Extracting Molecular Binding Relationships from Biomedical Text. *Proceedings of the Sixth Conference on Applied Natural Language Processing*. Seattle, Washington, April 2000.

Rus, V., Moldovan, D., and Bolohan, O. (2002). Bracketing Compound Nouns for Logic Form Derivation. *Proceedings of FLAIRS-2002*. Pensacola Beach, Florida, May 2002.

Shimbo, M. and Hara, K. (2007). A discriminative learning model for coordinate conjunctions. *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech Republic, June 2007.

Sparck-Jones, K. and Galliers, J. R. (1996). *Evaluating Natural Language Processing Systems: An Analysis and Review* No. 1083 in Lecture Notes on Artificial Intelligence. Berlin: Springer.

Srikumar, V., Reichart, R., Sammons, M., Rappoport, A., and Roth, D. (2008). Extraction of Entailed Semantic Relations Through Syntax-based Comma Resolution. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. Columbus, Ohio, June 2008.

Tjong, S. F. and Berry, D. M. (2008). Can rules of inferences resolve coordination ambiguity in natural language requirements specification? *Proceedings of the 11th Workshop on Requirements Engineering (WER'08)*. Barcelona, Spain, September 2008.