Information Extraction and Lexicon-Grammar

Patrick Watrin*
Centre for Natural Language Processing
1, place Blaise Pascal
1348 Louvain-la-Neuve, Belgium
watrin@tedm.ucl.ac.be

ABSTRACT

Information Extraction such as presented within Message Understanding Conference competitions consists, from some documents, in filling in template fields pre-established and organized around a precise scenario. With this aim in view, many researchers define patterns that are likely to target the information to extract. The aim of this paper is to show that syntactic databases of lexicon-grammar can be adapted to the context of the information extraction. Moreover, we view the conditions and the means to integrate them in an extraction system.

Keywords

Information Extraction, Lexicon-Grammar, Local Grammars, Finite-State Automata.

1. INTRODUCTION

Information Extraction, like most disciplines in Natural Language Processing, works differently depending on the preferred approach, i.e. linguistical or statistical. The linguistic approach, also called *knowledge-based* extraction, is based on grammar development. This grammar development inevitably implies the combination of an introspective approach and a corpus-based methodology, namely a specialized corpus. Indeed, the linguistic hypothesis is efficient if the underlying grammars exhaustively cover the target domain. The statistical approach falls within a more automatic process of resources learning. However, this learning requires the use of a corpus that has been (manually) pre-annotated according to the information to extract.

The last few years have shown a clear priority given to statistical methods, certainly because of their visible automation (even if linguistic methods have done well at the MUC). As for us, we believe that the orientation choice should be made according to the available resources. Roughly, a statistical approach is preferred if we have an annotated corpus. By contrast, a knowledge-based approach is chosen if we have lexical resources, and the knowledge of a linguist at our disposal, as it leads to a better general performance when the grammatical coverage is maximal.

The debate complexity is certainly not limited to these few lines which however enable us to more precisely root our research in the field of information extraction. Although we are by no means ruling out the pertinence of statistical approaches, our research is aimed at exploring the possibilities offered by the syntactic databases of the lexicon-grammar. It is thus for us a matter of proposing a linguistic solution to the problems raised by the information extraction.

Beyond this methodological choice, extraction procedures are highly similar. In [2] and [11], Grishman and Poibeau combine them with two consecutive levels of analysis: the *local analysis* that extracts relevant sequences according to the template fields and the *discourse analysis* that integrates the results. As will be seen, the lexicon-grammar particularly lends itself to the local analysis and more precisely to *syntactic* and *domain* analyses. As a reminder, the syntactic analysis aims at identifying heads of phrases or all the sentence constituents whether we choose for a partial or deep parsing. By contrast, the domain analysis identifies relevant events for a given scenario and their relations.

2. LEXICON-GRAMMAR

Grammarians have long apprehended syntax as an independent level of abstraction. Syntactic structures were seen as a combination of variables in which any entry could be inserted. However attractive, this view is much too simple as it neglects the idea of a lexical restriction.

The impossible dissociation between syntax and lexicology has led Maurice Gross to formalize language from the lexicon. This lexicon-grammar has, as an axiom, the submission of the linguistic abstraction to the scientific rigour. It results in the giving up of prescriptive methods in favour of descriptive ones and the creation of a theoretical model that describes language in a systematic way. It is the reason why this theory particularly lends itself to Natural Language Processing.

*This research is part of a PhD thesis financed by the Scientific Research Fund of the UCL (FSR).

DIR 2003 4th Dutch-Belgian Information Retrieval Workshop, Amsterdam, The Netherlands

AEGON WANTS TO FOCUS ON LIFE INSURANCE

The German insurance company Aegon has announced that it will sell the credit activities of its American subsidiary, Transamerica Finance, to the industrial conglomerate General Electric for 5.4 billion dollars.

In so doing, the group intends to disinvest its subsidiary which is in compliance with its reorientation strategy to focus on its core business i.e. life insurance.

05/08/03 10:07

Seller	NAME	Aegon
	DESCRIPTION	German insurance
Buyer	NAME	General Electric
	DESCRIPTION	industrial conglomerate
Transaction	OBJET	Transamerica Finance
	DESCRIPTION	American subsidiary
	DETAILS	credit activities
Amount		5.4 billion dollars

Table 1: Extraction Template

Moreover, the future of the lexicon-grammar has been bound to computer science since its beginning. Since [4], we consider the potentiality of a new type of syntactic parser[1] of which the theoretical root is precisely that of the lexicon-grammar and his formalism.

2.1 Syntactic Analysis

The linguistic data studied in the frame of lexicon-grammar is presented as a binary matrix (cf. Table 2) in which each line represents a predicate. This predicate is shown in an elementary sentence - which is the minimal unit of meaning - in order to disambiguate it. Each column of the matrix is a syntactic feature. The line-column intersections are filled in with a '+' symbol if the current entry validates the syntactic or semantic feature or with a '−' symbol if it does not.

In this way, if the syntactic analysis usually consists in recognising phrases step by step - from left to right or from right to left - the conceivable parsing procedure of a lexicon-grammar is different. The identification of a given sentence predicate allows us, with a simple check in the dictionary linked to the tables, to build up all the possible constructions for this predicate. The parsing amounts to the identification of the real context.

This analysis then combines the finite-state automaton that formalizes all the constructions to a given entry in the lexicon-grammar. We cannot pretend to realize this work manually as the verb tables themselves amount to more than 15,000 entries. In [14] and [10], we touch on this methodology of the *automates patrons*. As said earlier, the entire matrix is organized around an elementary sentence and certain number of distributional and transformational features. For each table of the matrix, we may then consider creating a reference automaton which would check all its features. The conversion of a given entry in its own automaton would then consist in removing the unchecked paths for this entry from the reference automaton.

[1]Although this parser does not exist yet, a lot of works are devoted to its development (cf. [14] et [10]).

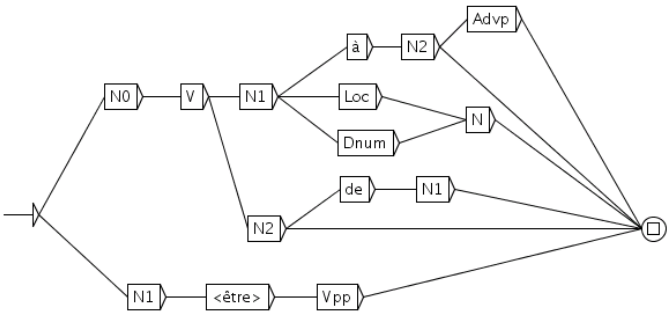

Figure 1: Partial automaton of table 36DT

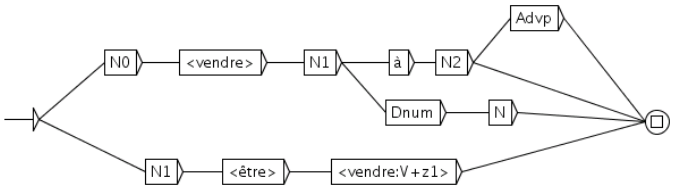

Figure 2: The automaton of *vendre*

2.2 Domain Analysis

Traditional extraction separates domain analysis from syntactic analysis. It results in a twofold procedure: syntactic analysis, applied to the text as a whole, identifies the heads of the different phrases; this information is then used by the domain analysis in order to identify, if need be, the presence of an extraction pattern. These patterns are often limited to simple nominal or verbal phrases and are not part of a sentence structure.

By contrast, the choice of the elementary sentence as a minimal unit allows the unification of these two analyses in one single process. Indeed, in addition to distributional as well as transformational syntactic indications, the tables of the lexicon-grammar present some semantic information. Thus, table 36DT (cf. Table 2) of the lexicon-grammar specifies two particularly relevant criteria in the context of company takeover:

Table 36DT

$N_0 = Nhum$	$N_0 =: V - n$	N_2 *bénéficiaire*		$N_1 = Nhum$	N_0VN_0pc	$N_1 = N - hum$	$N_1 =: V - n$	$N_1 = DnumN$	$N_1 =: coup$	*Prép* $=:$ *à*	*Prép* $=:$ *de*	$N_2 = N - hum$	*Ppv* $=:$ *lui*	*Ppv* $=:$ *y*	$N_2 = V - n$	$N_2 =: Nplobl$	$N_1 estVpp$	$N_1 =: Nabst$	N_0VN_1DnumN	$N_0VN_2(deN_1)$	N_0VN_1LocN	N_0VN_1 *à* N_2Advp
+	-	-	accepter	-	-	+	-	+	-	-	+	-	-	-	-	-	-	+	-	-	-	-
+	-	+	acheter	-	-	+	+	-	-	+	-	-	-	-	-	-	-	+	+	-	-	+
+	-	-	acquérir	-	-	+	+	-	-	-	+	+	+	-	-	-	-	+	+	-	-	+
+	-	+	céder	-	+	+	-	-	-	+	-	-	-	-	-	-	-	-	+	-	-	+
+	-	+	débourser	-	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-
+	-	+	mettre	-	-	+	-	-	+	+	-	+	+	-	-	-	-	-	-	-	+	-
+	-	+	offrir	+	+	+	+	+	+	-	-	-	-	-	-	-	+	-	-	-	-	-
+	-	+	payer	+	-	+	-	-	-	+	-	-	-	-	-	-	+	-	+	-	-	-
+	-	+	payer	+	-	+	-	+	-	+	-	-	-	-	-	-	+	-	-	+	-	-
+	-	-	prendre	+	-	+	+	+	-	+	-	+	+	-	-	-	-	+	-	-	+	+
+	-	-	racheter	+	-	-	-	-	-	+	-	-	-	-	-	-	-	-	+	-	-	+
+	-	-	racheter	-	-	+	-	-	-	+	+	-	-	-	-	-	-	-	+	-	-	+
+	-	-	rafler	+	-	+	-	+	-	+	-	-	-	-	-	-	-	+	-	-	+	-
+	-	+	rendre	+	-	+	-	+	+	+	-	+	+	-	-	-	-	+	-	-	-	-
+	-	-	reprendre	+	-	+	-	-	-	+	-	-	-	-	-	-	-	-	+	-	-	+
+	-	+	revendre	+	-	+	-	-	-	+	-	-	-	-	-	-	-	-	+	-	-	+
+	-	+	vendre	+	-	+	-	-	-	+	-	-	-	-	-	-	-	+	+	-	-	+

Table 2: Extract of table 36DT

- N_2 *beneficiary*: verbs of this class are said to be dative in the sense that they imply the exchange of the direct object N_1 of the elementary sentence N_0VN_1 *à* N_2, between the two human arguments N_0 and N_2.

- N_0VN_1DnumN and N_0VN_1Advp *à* N_2: although the information does not clearly appear in the table, the two complements - $DnumN$ and $Advp$ - that modify the elementary sentence within these two structures specify the amount of money at work in the exchange mentioned previously.

In a generic context, we cannot pretend to further precise the semantic criteria (type of beneficiary : individual, entreprise,... ; type of exchange : stock market transaction,...). Indeed, the syntactico-semantic analysis that derives from the use of such a formalism fits any situation of enunciation. However, if we restrict the analysis to a particular context (e.g. company takeover), we are then able to specify the semantic characteristics. Moreover, this idea reminds the Harrissian notion of sublanguage defined in [5] and taken over by [16] and [3] among others in a theoretical frame which is similar to ours.

This concept especially applies to the extraction of information that focuses on a closed and predefined set of events. From then on, we may consider restricting the semantic focus of essential complements, which allows us to establish more accurate and consequently more efficient templates (cf. Table 3).

3. FIRST EMPIRICAL RESULTS

Let us remind that the extraction process proposed here carries out the identification, within a text, of the various fields of a predefined form. Our assumption is that each of these fields corresponds to the argument of an elementary sentence (i.e. the significance unit of the lexicon-grammar).

We are basing this analysis on a corpus of 1,000 financial dispatches. This corpus includes 430,000 words distributed among 7,893 sentences (titles included). The extraction proper is organised around the scenario mentioned before (cf. Table 1) and a toy lexicon-grammar of eight verbs derived from table 36DT only (cf. Table 3).

3.1 Used Metrics

- RECALL : ratio of the number of relevant extracted documents to the total number of documents. The term *document* should here be understood as any sentence having as predicate one of the eight verbs of our toy lexicon-grammar. It follows that a relevant document is a sentence of which the elementary format is N_0 V N_1 *à* N_2, i.e. the standard format of table 36DT.

- PRECISION : ratio of the number of relevant extracted documents to the total number of extracted documents.

- F-MEASURE : measurement combining the above two in order to balance them.

$$F = \frac{2PR}{P+R}$$

3.2 Examples and results

PATTERNS					
N_0	V	N_1	N_2	*Dnum*	*Advp*
<buyer>	***acheter***	<transaction>	<seller>	<amount>	<amount>
<buyer>	***acquérir***	<transaction>	<seller>	<amount>	<amount>
<seller>	***céder***	<transaction>	<buyer>	<amount>	<amount>
<seller>	***offrir***	<transaction>	<buyer>	-	-
<buyer>	***racheter***	<transaction>	<seller>	<amount>	<amount>
<buyer>	***reprendre***	<transaction>	<seller>	<amount>	-
<seller>	***revendre***	<transaction>	<buyer>	<amount>	-
<seller>	***vendre***	<transaction>	<buyer>	<amount>	-

Table 3: Example of *financial* patterns

	RECALL (%)	PRECISION (%)	F-MEASURE
acheter	66.7	100	80
acquérir	24.2	100	53,4
céder	56.5	100	72,2
offrir	100	100	100
racheter	75	100	85,7
reprendre	0	-	0
revendre	100	100	100
vendre	85.7	92.3	88,9
	63.5%	**98.9%**	**77,3**

3.2.1 Successful extractions

1. Idenix : <BUYER>le groupe pharmaceutique suisse Novartis</BUYER> a acheté <AMOUNT>pour 582 millions de dollars</AMOUNT> <TRANSACTION> 51% de la société américaine de biotechnologies, spécialisée dans les produits contre l'hépatite B et les antiviraux</TRANSACTION>.
   Idenix : the Swiss pharmaceutical group Novartis has purchased for 582 million dollars 51% of the American biotechnological company specialized in drugs against hepatitis B and in antiviruses.

2. <SELLER>Rexel, la filiale de distribution de matériel électrique de PPR,</SELLER> a cédé <TRANSACTION>Gardiner Group, Stentorius et JLD, sociétés spécialisées dans la distribution de matériel de sécurité électronique,</TRANSACTION> <BUYER>à Electra Partners</BUYER> <AMOUNT> pour 112 millions d'euros</AMOUNT>.
   Rexel, the electronics distribution subsidiary of PPR, has sold Gardiner Group, Stentorius and JLD, companies specialized in the distribution of electronic security equipment, to Electra Partners for 112 million euros.

3. <SELLER>Le groupe allemand d'énergie et de services aux collectivités E.ON</SELLER> prévoit de vendre <TRANSACTION>Viterra Energy Services </TRANSACTION> <BUYER>à CVC Capital Partners</BUYER> <AMOUNT>pour un montant d'environ 1 milliard d'euros</AMOUNT>, selon des agences de presse.
   The German energy and community service group E.ON is planning to sell Viterra Energy Services to CVC Capital Partners for an amount of approximately 1 billion euros, according to press agencies.

3.2.2 Aborted extractions

4. La banque au lion a annoncé ce matin qu'elle allait racheter **40% de la participation d'Aventis dans Rhodia**, soit 9,9% du chimiste.
   The 'lion' bank announced this morning that it is going to buy back 40% of the Aventis participation in Rhodia, i.e. 9.9% of the chemical company.

5. *La banque au lion a annoncé ce matin qu'elle allait racheter **40% de la participation d'Aventis dans Rhodia *à Aventis***, soit 9,9% du chimiste.
   The 'lion' bank announced this morning that it is going to back buy 40% of Aventis participation in Rhodia to Aventis, i.e. 9.9% of the chemical company.

6. Pour justifier ses griefs, Valauret met en lumière la situation de Rhodia qui **"s'assombrit tous les jours alors que l'actionnaire majoritaire Aventis va céder sa participation de 25,2%"** abandonnant Rhodia, ses salariés et ses actionnaires, après avoir cautionné la stratégie de ces cinq dernières années.
   To justify its grievances, Valauret highlights the situation of Rhodia which grows dark every day whilst Aventis, the majority shareholder, is going to sell its 25.2% participation abandoning Rhodia, its employees and shareholders after having supported the strategy of these last five years.

7. Pour consolider son pouvoir, Racamier fait venir **Bernard Arnault, un inconnu qui avait repris Boussac (propriétaire de Christian Dior)**, désireux de rapprocher la prestigieuse marque de Christian Dior Parfums (licence détenue par LVMH).
   To consolidate his power, Racamier is bringing in Bernard Arnault, a stranger who had taken over Boussac (owner of Christian Dior), anxious to bring closer the prestigious Christian Dior Parfums brand (licence owned by LVMH).

8. Alors que la loi fédérale américaine contraint les groupes pharmaceutiques à **vendre leurs médicaments** au meilleur prix offert aux clients privés, Bayer et GSK auraient changé le nom de plusieurs de leurs produits avant de les vendre à un prix moins cher à un assureur privé.
   While US federal law forces pharmaceutical groups to sell their drugs to private customers at the best possible price, Bayer and GSK have reportedly changed the name of several of their products before selling them at a lower price to a private insurance company.

3.3 Recall problems

3.3.1 The constituents concatenation

Some sentences such as (4), although including the constituents N_0, N_1 and N_2 of the elementary sentence N_0 V N_1 *à* N_2, do not have the same syntactic structure. So, in our example, the N_2 beneficiary *à Aventis* no longer appears under its elementary form as indirect object but as object of N_1, subject of the transaction. This type of exception to the rule is very hard to explain from a transformational standpoint. We would have there for to presuppose the initial the-

oretical sentence (5) which, by deletion of the redundancy, would give the structure (4).

3.3.2 Quotations

The quotations appearing in the press are often incomplete. Sentence (6), which does not include the essential constituent N_2, will therefore not be considered.

3.3.3 Syntactic differentiation

As we have seen, the elementary sentence of the lexicon-grammar allows the semantic disambiguation of the predicate by placing it in a syntactic context. The essential constituents are thus as many elements allowing us, outside a real enunciation situation, to distinguish the different meanings of a predicate. These constituents however are not always essential from a syntactic standpoint and can, consequently, disappear thereby preventing the extraction (cf. (7)).

3.4 Precision problem

We have encountered only one single extraction example which does not fit in our scenario. This error is very easily explained. We have taken the option to work from an heterogeneous corpus which, although dealing with finance, largely exceeds the context we had set for ourselves. Sentence (8), because it includes the term *sell*, is naturally associated with our form.

4. FUTURE DEVELOPMENTS

The globally positive results of our first evaluation confirm our initial assumption and demonstrate that the elementary sentence lends itself particularly well to extraction procedures. The essential constituents associated to each predicate within the syntactic matrix are precisely those carrying the essential information of the sentence. However, the analysis of these results leads to a double observation.

- The constituents of the elementary sentence, essential to the semantic disambiguation of the predicate, are, from a purely syntactic viewpoint, not always essential. They can therefore occasionally disappear.
- The coverage resulting from our toy lexicon-grammar is far from optimal. Many dispatches pertaining to our scenario could not be extracted simply because they did not include any of the eight verbs.

An immediate solution to these problems would consist in an exhaustive study of the lexicon specific to the extraction domain. Such a study, aimed at creating a specialty lexicon-grammar, would as well allow claiming a maximal coverage as, for each predicate, targeting at best the truly essential constituents. It can however not be carried out manually. As a matter of fact, in view of the restricted semantics of a scenario, we could not consider defining many extraction domains.

Most pattern learning systems (cf. [7], [12], [13], [9], [17] and [15]) originate in a distributional analysis such as defined at an early stage in [6] and [1]. That analysis is based on the mere observation of the elements in enunciation situation and, therefore, lends itself to the automatism we are aiming at.

> *La technique fondamentale de recherche employée dans le présent travail est l'analyse distributionnelle. En outre, la méthode de substitution, l'analyse componentielle et certains éléments de l'analyse transformationnelle sont appliqués aux différentes étapes de la recherche.* [1]

It is the same principle of substitution, essential to the distributional analysis, which is implemented by the *mutual bootstrapping* developed in [12] and [13] in order to infer from new extraction dictionaries. It should be noted that the concept of mutual triggering off has something of the nature of a distributional synonymy also emphasised in [8]. So, we start from a number of predefined patterns (in the same way as in [15]) which we submit to a specialized corpus. What emerges from this analysis is the extraction of a lexicon of arguments. Submitting the combination of the arguments to the corpus will then enable us to obtain new predicates that will complete our library of extraction patterns.

Our approach cannot however be limited to the one in [12] and [13]. [7], [17] and [15] have before us highlighted the weakness of patterns inferred by *AutoSlog(-TS)* ([12]) (e.g. *killed <victim>*, *killed with <instrument>*, *to kill <victim>*, *<perp> attempt to kill*,...). On the one hand, the relations between constituents are not recognized (e.g. *<perp> kill <victim> = <perp> kill + kill <victim>*) and are rebuilt only afterwards. On the other hand, each pattern hinges on a fixed lexical anchor, typically a verb, which is not subject to generalization. One therefore observes a large redundancy of patterns (e.g. *<kill> <victim> = kill <victim> + to kill <victim> + killing <victim>*).

The syntactic formalism of the lexicon-grammar allows us to go over these limits. Once a new predicate has been identified, the syntactic analysis of its environment and, more particularly, of the arguments leading to its identification, will enable us to extract the elementary sentence (i.e. its entry in the lexicon-grammar). The same elementary sentence, increased by its associated transformations, will clear the redundancy mentioned earlier and, at the same time, will enable the extraction of the related constituents.

5. CONCLUSION

The objective we set for ourselves in this exploratory research was twofold. We wanted first to assess the relevance of the lexicon-grammar as a source of knowledge for information extraction and, second, to analyze the conditions of its integration within an extraction system.

The results we have achieved (63.5% recall and 98.9% precision) demonstrate that, in addition to the minimal significant unit of the lexicon-grammar, the elementary sentence

represents an informational unit and efficiently meets the extraction needs. These results however only reflect the extraction capacity of the eight predicates used. As stated in our introduction, a knowledge-based system is fully efficient if its grammar entirely covers the target domain. Therefore, before considering a robust system, we will necessarily have to implement a learning process such as mentioned before.

6. REFERENCES

[1] J. D. Apresjan. Analyse distributionnelle des significations et champs sémantiques structurés. *Langages*, 1:44–74, 1966.

[2] R. Grishman. *Information Extraction: Techniques and Challenges*, pages 10–27. 1997.

[3] R. Grishman. Adaptive information extraction and sublanguage analysis. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (IJCAI-2001)*, 2001.

[4] M. Gross. *Méthode en syntaxe : régime des constructions complétives*. Hermann, Paris, 1975.

[5] Z. S. Harris. *Structures mathématiques du langage*. Dunod, Paris, 1971.

[6] Z. S. Harris. Discourse analysis. *Language*, 28:1:1–30, 1974.

[7] S. B. Huffman. Learning information extraction patterns from examples. In *Proceedings of the Workshop on New Approaches to Learning for Natural Language Processing (IJCAI-1995)*, 1995.

[8] C. Leclère. Une approche syntaxique de la synonymie. *Cahiers de l'institut de linguistique de Louvain*, 2:77–94, 1999.

[9] S. S. . W. Lehnert. Learning domain-specific discourse rules for information extraction. In *Spring Symposium on Empirical Methods in Discourse Interpretation and Generation (AAAI-1995)*, 1995.

[10] S. Paumier. *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. PhD thesis, Université de Marne-la-Vallée, 2003.

[11] T. Poibeau. *Extraction d'information : du texte brut au web sémantique*. Hermès, Paris, 2003.

[12] E. Riloff. Automatically generating extraction patterns from untagged text. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-1996)*, pages 1044–1049, 1996.

[13] R. J. . E. Riloff. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-1999)*, pages 474–479, 1999.

[14] E. Roche. Une représentation par automate fini des textes et des propriétés transformationnelles des verbes. *LinguisticæInvestigationes*, 17:1:189–222, 1993.

[15] P. T. . S. H. Roman Yangarber, Ralph Grishman. Automatic acquisition of domain knowledge for information extraction. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, 2000.

[16] N. T. N. . N. Sager. *The computability of strings, transformations, and sublanguage*, pages 79–120. John Benjamins, Amsterdam & Philadelphia, 2002.

[17] D. F. . W. L. Stephen Soderland. Automatically learned vs. hand-crafted text analysis rules. In *CIIR Technical Report*, 1997.