Natural Language Engineering **3** (2/3): 171–190. Printed in the United Kingdom

© 1997 Cambridge University Press

TextNet – A text-based intelligent system

SANDA M. HARABAGIU
Parallel Knowledge Processing Laboratory, University of Southern California, Los Angeles, CA 90089-2562
e-mail: harabagi@usc.edu

DAN I. MOLDOVAN
Department of Computer Science and Engineering, Southern Methodist University, Dallas, TX 75275-0122, USA
e-mail: moldovan@seas.smu.edu

(*Received 31 December 1996; revised 3 April 1997*)

Abstract

A large collection of texts may be reached through the Internet and this provides a powerful platform from which common-sense knowledge may be gathered. This paper presents a system that contains a core knowledge base structured around WordNet, a lexical database, capable of extracting contextual information from a given input text. Such context information is then used to retrieve other texts from the Internet that relate to that context. When processed by the system, these new texts bring more information that represents an enhanced domain context for the initial text. This is an incremental method for text processing that acquires domain knowledge from other texts. The paper describes the system architecture, its core knowledge base and inference engine, and the acquisition of new knowledge from corpora.

1 Introduction

1.1 Knowledge representation challenges

Knowledge representation systems are of central importance to the field of AI since they provide the explicit declarative information upon which inference can be done. Knowledge representation systems have to tackle a variety of problems encompassing expressiveness power, quality and efficiency of possible inference, completeness, scalability, and correctness of ontological principles.

Realistic knowledge bases have to be very large, perhaps as large as millions of concepts. Unfortunately, due to the high interconnectivity of knowledge, it is not possible to partition a knowledge system into domains. The modularized approach that has been used successfully in other areas does not apply here.

Experience has shown that there are major problems with a domain scalability. Modest size systems working in restricted domains were successfully implemented, but they quickly failed when the domain changed slightly. Lenat (1995) and others have called this the 'brittleness' problem. A serious problem that hampers the scalability of knowledge representation systems is that reasoning is usually implemented

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

with lots of rules encoding simple facts like "if it rains, then the lawn is wet". Since knowledge is infinite, it will not be possible to generate a complete set of rules. In fact, the number of rules gets quickly out of hand even for systems with modest capabilities. Moreover, combining inference rules from different domains generates correctness problems.

These are only but a few problems in knowledge representation systems. Then, the question is what knowledge representation method gives us access to unlimited, scalable information? In this paper we argue that the answer to this increasingly important question is *natural language*.

1.2 Text-based knowledge representation

Sowa (1984), Ali and Shapiro (1993), Iwanska (1992) and Schubert and Hwang (1989) were among the first to suggest the use of natural language to represent knowledge. A knowledge representation system which uses textual information – coined by Jacobs (1992) – Text-Based Intelligent Systems (TBIS), may work like this: a text that contains domain information is transformed by a natural language processing system into a knowledge representation. The sophistication of the natural language processing may vary from simple tagging to 'deep understanding', and this determines the power of such a system. Then, the user or the system provides information also in the form of natural language that needs to be transformed by the same natural language processing system in a form of knowledge. Then, the text and user representations are compared, and inferential information is extracted.

Most likely, the TBIS approach still uses a knowledge base; however, the difference is that here the knowledge base is used only to interpret the text, and the text provides domain and context information. Thus, the knowledge base does not have to encode unnecessary facts and procedures.

We believe that with some core knowledge base the textual context may be depicted, which in turn may bring into focus many other texts from a large corpora that will further enhance the domain knowledge of the original text. This is an incremental approach to knowledge acquisition and text understanding.

Knowledge shared, taught or expressed in natural language is considered to have by far the best expressiveness, although it is very ambiguous. Other representation formalisms eliminate the ambiguity aspect, but limit the wealth of knowledge communicated. Moreover, a considerable effort is needed to translate into another representation scheme the immensity of facts and descriptions stated in the overflow of textual information characteristic to modern life.

The use of a knowledge representation scheme that incorporates natural language texts implies a form of natural language understanding and inference. In reality because of the incremental processing used, shallow text processing methods seem adequate.

The development of Internet clients capable of supporting aspects of domain-independent textual inference becomes more and more desirable. Today the web is populated by several softbots (software robots as coined by Oren Etzioni (1996)) among the panoply of primitive, but popular information services. The existent

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

crawlers available for softbots (Kirk *et al.* 1995; Shakes *et al.* 1996) tend to rely on heavy indexing instead of recent developments from information retrieval (Broglio *et al.* 1993; Liddy and Myaeng 1993).

The TBIS solution is appealing since it replaces the construction of a large knowledge base with the acquisition of knowledge via text understanding. The volume of machine readable text is growing exponentially. On Internet, information in a form of a text is available in real-time on almost every subject. Newspapers, journals, books, etc. are becoming available electronically more and more.

1.3 Proposed solution

This paper investigates a technique for building network-based information brokers, capable of (1) retrieving texts that display contextual closeness to a given text, and (2) interpreting the gathered information in coherence with the source text. The contextual information brokers are components of TextNet Harabagiu and Moldovan 1996*c*, a text-based knowledge representation system capable of inferring commonsense knowledge from texts gathered in a corpora or spread over the web.

The interpretation of several texts as having similar contexts is determined by the evaluation of a coherence metric that captures the common concepts and lexical relations spanning any pair of texts. This measure explains the observation that a pair of texts using the largest number of common concepts in the same semantic constructs tend to have closer contexts.

In our system we use WordNet 1.5 Miller 1995 as the core linguistic knowledge base, since it has a broad semantic coverage. Bridging the glossary definitions from WordNet with textual information generates a complex knowledge infrastructure.

TextNet relies on the simple but consistent semantic principles of WordNet, and it extends its micro contextual concept definitions to complex structures, capable of dealing with a large variety on compound concepts, often encountered in texts. The 14 semantic relations encoded in WordNet, as will be seen in section 2.2, apply broadly throughout English and are assumed to be contextually independent. Moreover, TextNet, through its inference engine, addresses some of the WordNet limitations, like the ability to represent goals, plans, action preconditions and results. A query about a concept is transformed into a unification search of the instances of that concept on the web of texts, producing a defining context revealed by the logical connections to other typical concepts.

2 System architecture

2.1 TextNet organization

Figure 1 gives a functional representation of our system. The knowledge sources consist of on-line dictionaries, extended WordNet and Internet text documents. The components of the linguistic knowledge base are shown in connection with the components of the run-time system capable of processing texts. The result of text inference is the complex knowledge structures illustrated on the right of the figure,

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

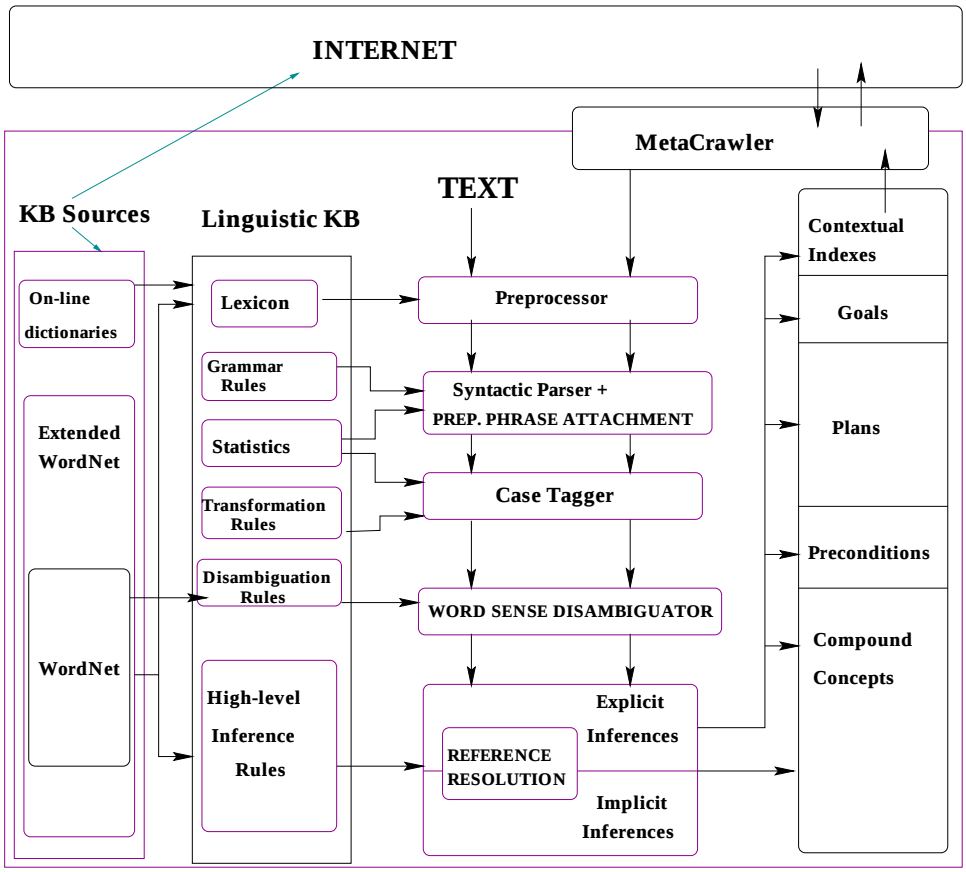


Fig. 1. TextNet organization.

containing compound concepts, preconditions, plans and their recognized goals and contextual information connecting the processed text with other texts from the the Internet.

TextNet has been partially implemented on the SNAP parallel machine Moldovan *et al.* 1992, using the Marker Propagation Networks computational model Lin *et al.* 1994. The inference system is based on the codification of complex propagation rules, implemented as C lightweight processes Harabagiu and Moldovan 1997*b*. On average,the processing of a sentence required 25 propagation rules and the speedup obtained in the magnitude of hundreds proves the requirements of parallelism when processing semantic networks encoding millions of relations.

The linguistic processing part of TextNet was developed as an information extraction system for the Message Understanding Conference Sundheim 1993. The extensions of WordNet as well as the high-level inference module were developed as a means to evaluate text coherence. TextNet emerged as a solution to the limitation we found in WordNet Harabagiu and Moldovan 1997*a*. The semantic information provided by WordNet is used in many forms of linguistic inference. Prepositional

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

Table 1. *The number of words and concepts in WordNet 1.5*

Part of speech	Words	Concepts
noun	107,484	60,5557
verb	25,768	11,363
adjective	28,762	16,428
adverb	6,203	3,243
Total	168,217	91,591

attachments Harabagiu 1996 and case tagging integrate WordNet information. Text implicatures are recovered by the path-finding algorithm using the extended WordNet.

2.2 Constructing a knowledge base around WordNet

WordNet developed at Princeton by a group led by George Miller Miller 1995 is a semantic dictionary in which words are searched based on conceptual affinity with other words. WordNet 1.5 is publicly available and a Java interface is reachable at `http://www.cogsci.princeton.edu/~wn/w3wn.html`. It covers the vast majority of nouns, verbs, adjectives and adverbs from the English language. The words in WordNet are organized in synonym sets, called *synsets*. Each synset represents a concept. There is a rich set of relation links between words and other words, between words and synsets, and between synsets.

The current version, WordNet 1.5 contains a large network of 168,217 words, organized in 91,595 synsets, connected through 345,264 semantic links. The noun concepts are spanned by 11 *isa* hierarchies whereas the verb concepts are classified into 558 *isa* hierarchies. Taxonomies of *has_part*, *has_member* or *is_stuff_of* relations exist also between noun concepts. Causality between actions is revealed by two semantic relations between verb concepts: the *entail* and the *cause_to* relations. A negation lexical operator is implemented with the *antonym* relation, linking words with opposite meanings. Adjective concepts are clustered by the *similarity* semantic relations. There are also *attribute* relations, linking semantically adjective concepts to noun concepts. Words are also linked by morphologically derived relations: *pertainym* or *past_participle*.

In WordNet, a major source of information is provided by the glosses attached to almost every synset. WordNet glossaries, viewed as dictionary definitions, induce supplementary information. Disambiguated glosses, transformed in network representations can increase significantly the connectivity of the lexical knowledge base. Moreover, topological relations between synsets and the definitions in which they participate are possible. These are plausible indexes for their functional contexts. Conversely, each definition may also be considered as a constructing block of a local context outlined by that concept.

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

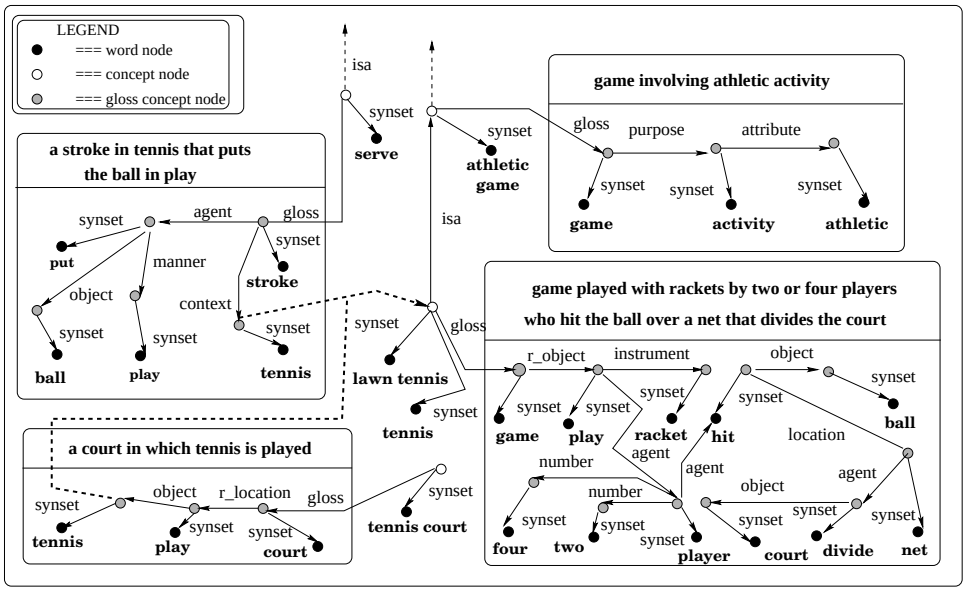


Fig. 2. Contextual pointers in WordNet derived by concept `tennis`.

As an example, in figure 2 are illustrated several of the 36 topological relations spanning from the WordNet concept {`tennis, lawn tennis`}. A multitangled knowledge base is built this way. The concept is linked to all the concepts from its own gloss as well as to those from the glosses of its hypernyms, for example the concept {`athletic game`}. It is also related to concepts like {`serve`}, {`tennis court`} (and their hypernyms and related gloss concepts) because concept {`tennis, lawn tennis`} is a disambiguated gloss concept with defining roles for these concepts. As we can see, a gloss concept relates topologically to many other gloss concepts. In figure 2 concepts {`serve`} and {`tennis court`} are topologically related, due to the contextual pointers starting at the gloss concept {`tennis, lawn tennis`} that is part of their gloss, and addressing the regular WordNet concept {`tennis, lawn tennis`}.

Topological relations contribute to the lexical relations recognized in the gloss transformation as well as to the regular semantic relations from WordNet in order to model textual context, as shall be clarified by the contextual inference models presented later.

Missing from WordNet is knowledge about preconditions, plans and their corresponding goals that may apply in some situation. This is normal since WordNet was intended to be only a lexical database. Procedural and precondition knowledge specific to a situation must be gathered from other texts. This is acquired by the TextNet inference engine.

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

3 Inference engine

3.1 Path-finding

Whenever TextNet is presented with a new text, implicit inferences are sought by combining WordNet and corpora information. The text is disambiguated by identifying first the explicit, and then part of the implicit connections between the concepts. The relations provided by transformational grammars are integrated with semantic connections involving defining features and standard WordNet relations. More complex inferences are drawn only after the semantic connections are recognized. Semantic paths linking pairs of concepts are the means through which logical connections are identified. The paths are searched using a set of primitive inference rules, implemented as constrained chains of WordNet relations. The relations that are asymmetric and transitive are more useful for inference than others since often they support deductions. For example, going up on isa hierarchy of nouns or verbs, takes us from more specific to more general concepts and hierarchical deductions are possible. The same for some of the verb relations such as entail or causation. For example, if a man is snoring, than he is sleeping, and further he is temporarily unconscious.

However, inferences sometimes need to go against these transitive relations, and conclusions are not precise. For example, if one is sleeping, we don't know for sure that one is snoring, but is plausible. In these cases, TextNet performs abductions, providing the best explanation. It adds new information provided by retrieved texts with close contexts, and enforces or discards the abduction.

A large number of inference primitives are implemented as propagation rules guiding markers in the extended WordNet semantic knowledge base. Combination of pairs of WordNet relations like *isa* and *entail* or *is_part* and *isa* constitute the inference primitives. Markers are placed on nodes, and their propagations connect the visited nodes. When markers are created are filled with propagation rules that enable them to reach all network concepts semantically related. A typical marker format contains fields that indicate the sentence and the word where that marker originated, marker propagation function, the processing functions called by the marker and the markers with which it can collide.

After markers are placed on nodes, they start to propagate according to their propagation rule. New markers are spawned every time a node contains more than one outgoing link that is part of its propagation rule. Whenever a marker reaches a node that does not have an outgoing relation as part of its propagation rule, the marker stops. The propagation trail of a marker becomes part of that marker history, thus markers become fatter as they propagate.

An example of a propagation rule for markers placed on verbs is any combination of relations `isa`, `caused_to`, `entail` and `reverse_entail`. Markers continue to propagate as long as the propagation rule allows. The algorithm detects and avoids cycles by simply checking whether or not a marker has visited already a node.

After all marker propagations cease, the collisions that took place during marker propagations are detected. To each marker collision it corresponds a path. The path is found simply by inspecting the path history of the colliding markers and linking

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

the two halves. Each marker participating in a collision provides half of the path as it appears in its field called path.

The assumption that every path recovers from the knowledge base an explanation of the lexical relation connecting the ends of the path in the text differentiates this inference method from other link climbing approaches or from Sowa's conceptual dependency theory Sowa 1984. The markers are intelligently focused to reach knowledge base nodes connected by the same lexical relations as their starting nodes. Their propagation follows in fact a trail of case constraint propagations along the knowledge base relations. Marker propagation rules, operating in process flow mode are most different from marker passing techniques, because their collected results reflect smart searching strategies.

3.2 Inference extraction

After the semantic paths spanning a text are detected, complex inferences can be generated. The chains of deductions that can be extracted from a sentence provide two types of inference: (1) inference providing supplementary information pertaining to a sentence, and (2) abductions explaining the relations between sentences.

The central idea for inference extraction is to substitute a sentence concept that is the source of a relation with the destination of that relation. If it was a gloss relation, then the whole gloss fragment found in the context is used for the substitution. For example, we know that `snore`→*entails*→`sleep`. Based of this relation, from sentence `The guy next to me snored like a horse` we conclude that `The guy next to me slept`. Note that the inferences drawn are subject to linguistic constraints.

Lexical entailment between verbs in WordNet is implemented with the assumption of reference to states that involve logical implication between the propositions using the two verbs. Verbal troponymy, implemented through *isa* relations encode manners of performing the same action, and does not refer to any logical state or proposition. Therefore, there is an intrinsic distinction between reverse entailments and subsumptions. The next section offers more details and examples.

4 Acquisition of contextual knowledge

4.1 Intra-sentential local context

In an extensive knowledge base like ours, it is typical to have a very large number of possible connections between two concepts Harabagiu and Moldovan 1996*a*. However, only some of these connections are relevant for a text, while other connections may be relevant for another text. The recognition of the local textual context is done by finding the connections from the knowledge base that explain the lexical relations from the text graphs. This is accomplished by recognizing the set of knowledge base connections corresponding to each pair of concepts from a text graph linked by a lexical relation.

Figure 3 illustrates some of the nodes and relations from the knowledge base that are related to the text graph of sentence `S1`, defining its local context.

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

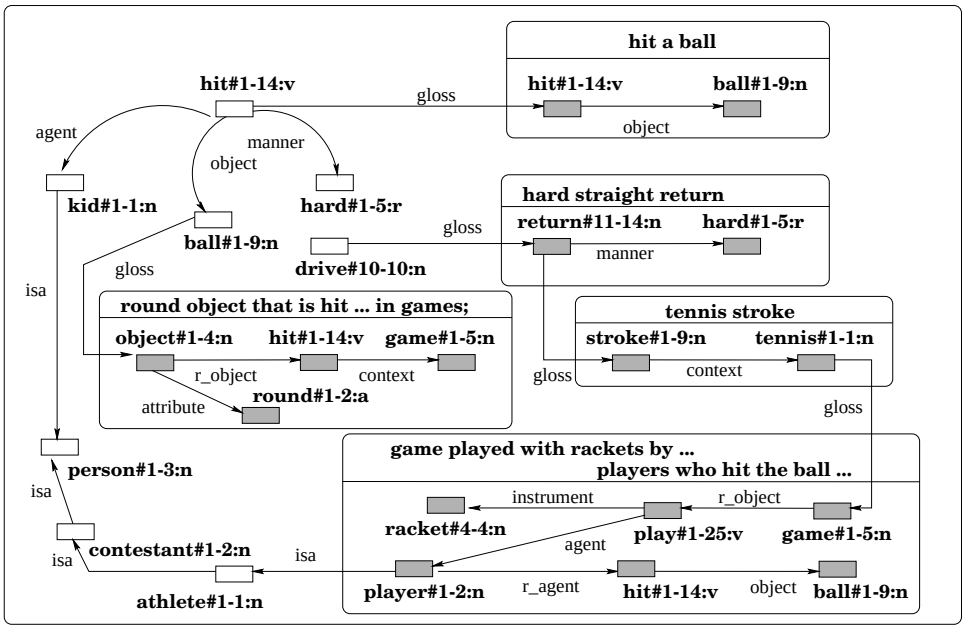


Fig. 3. Local connections for sentence S1.

```
S1: The kid hit the ball very hard.
```

In figure 3 the notation *word*#*i*—*j*:POS represents the concept corresponding to sense *i* of *word* out of a total of *j* senses in WordNet1.5 for the part-of-speech listed in the POS field. The WordNet notation for part-of-speech is n for nouns, v for verbs, a for adjectives and r for adverbs. For example hit#1-14:v means sense # 1 of verb hit which has 14 senses. In WordNet, the senses of a word are ordered according with the frequency of their use, the smaller the number *i*, the more frequently used. In figure 3 a rectangle represents a concept node and all its corresponding word nodes and *synset* links.

The local knowledge base connections are grouped around their respective lexical relations from the text graph. We illustrate in Table 2 the inference chains and the contexts for each of the three lexical relations from S1.

The links and concepts from these connections define the local context for sentence S1. They provide with a variety of new plausible information (e.g. the kid hit the ball with a racket; he played tennis on a tennis court, his return might have been a drive, etc.). Local knowledge base connections offer a semantic interpretation of each lexical relation from the text.

4.2 Inter-sentential global context

The next step is to take advantage of the collective meaning of all sentences in the text by finding connections between the local contexts. Inter-sentential connections

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

Table 2. *Chains of inference and local context*

Sentence	Inference source
Inference chain for relation agent(kid, hit)	
The kid hit the ball	hit (in gloss of ball)→
The kid hit the ball in the context of a game	*context*→ game game same as in gloss of tennis
The kid hit the ball in the context of tennis	tennis played by players
The kid hit the ball in the context of tennis, a game played by players	
The kid = player hit the ball	unification
The player hit the ball	player→*isa*→ athlete
The athlete hit the ball	athlete→*isa*→ contestant
The competitor hit the ball	contestant →*isa*→ person
The person hit the ball	kid →*isa*→ person
The kid hit the ball	
Inference chain for relation object(hit, ball)	
The kid hit the ball	hit (in gloss of ball)→ *context*→ game
The kid hit the ball in a context of a game	game same as in gloss of tennis
The kid hit the ball in a context of tennis	tennis → *gloss*→ a game played by...
... a game played by players who hit the ball	
Inference chain for relation manner(hit, hard)	
... hit the ball hard	hard (in gloss of drive → *r_manner* → return
... returned the ball hard	return → *gloss*→ stroke
... stroke the ball hard	stroke → *context*→ tennis
... in a context of tennis	tennis → *gloss*→ game
... in a context of a game	game → *r_object*→ play
... game played	play → *agent*→ player
... game played by players	player → *r_agent*→ hit
... who hit the ball	

are sequences of knowledge base relations and nodes linking intra-sentential connections and providing the global context of a text. Let us consider the text:

S1: The kid hit the ball very hard.
S2: It landed almost always near the baseline.

Figure 4 illustrates some of the nodes and relations from the knowledge base that establish connections between sentences S1 and S2. Two of these text connections are shown in Table 3.

The two textual connections perform two functions: (1) they reinforce local context (e.g. by confirming that the kid was playing tennis, derived from the second textual

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

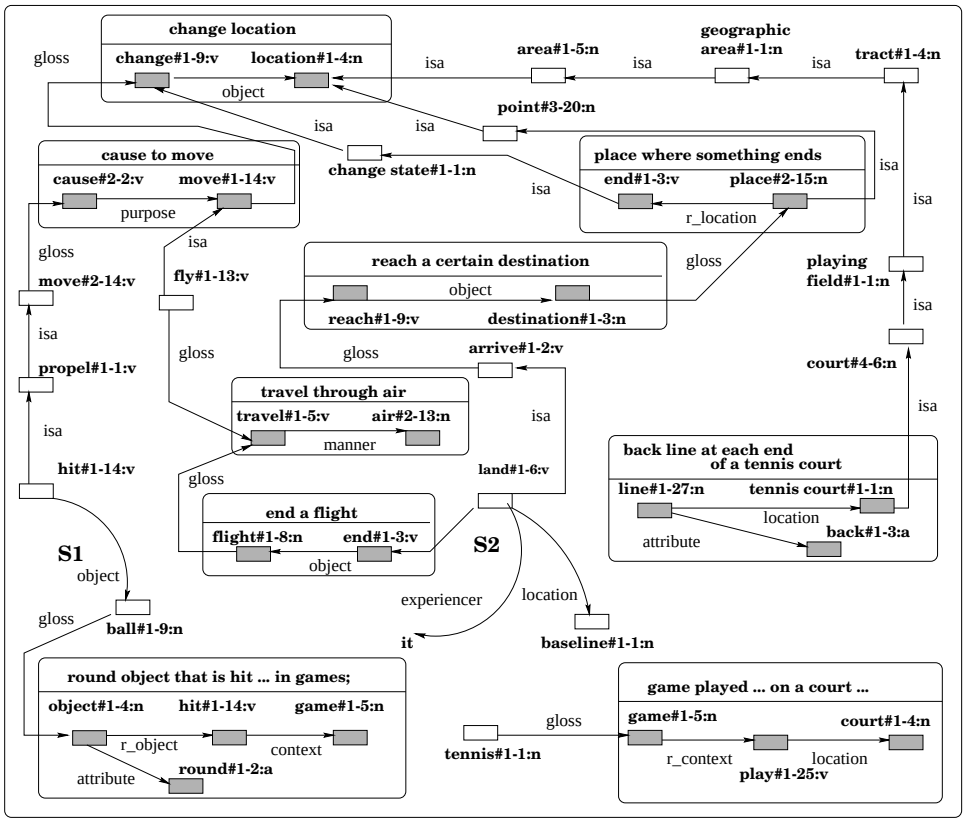


Fig. 4. Connections between sentences S1 and S2.

connection), and (2) provide with new contextual information (e.g. the ball travels through the air in a tennis game).

Figure 4 also illustrates interesting contextual information that produces inferences of the type: (1) the ball is hit with the purpose to have it moved to another location, and (2) an object that has been traveling through the air arrives to its destination when the flight ends.

These and many other inferences can be extracted from the net of concepts and links contained in the global context of the considered text.

5 Acquisition of additional knowledge from corpora

5.1 Context brokers

The contextual representation provided by the inference system does not contain all the information implied by the text, due in part to the limitations of world knowledge covered by WordNet and its extension. The role of contextual brokers is to gather from the Internet world knowledge related to the context of a text, in order to provide with additional information that may motivate the intentional structure of a text, among other things.

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

Table 3. *Chains of inference and global context between S1 and S2*

Sentence	Inference source
Context between hit *and* land	
The kid hit the ball very hard	hit →*isa*→ propel
... propelled the ball ...	propel →*isa*→ move
... move the ball ...	move →*isa*→ change location
... the ball changed location..	object of change location is ball
the ball = it reached location ...	unification, place→*isa*→ location
It reached a place where	destination→*gloss*→ where
something ends..	something ends
It reached destination...	arrive→*gloss*→ destination
It arrived...	land →*isa*→ arrive
It landed far away	
Context between hit *and* baseline	
The kid hit the ball far away	hit (in gloss of ball)→ *context*→ game
... hit the ball in the context of a game	game same as in gloss of tennis
... hit the ball in the context of tennis	tennis → *gloss*→ ... on court
... tennis played on tennis court	baseline → *gloss*→ line
... baseline is a line on tennis court..	on tennis court

A way of gathering contextually close world knowledge is to retrieve documents involving some of the concepts relevant to the context, and then to measure the coherence between the returned information and that provided by the initial text.

Currently there are several Web search services, such as Galaxy, Infoseek Lycos, Open Text, WebCrawler, Yahoo , Savvy Search and MetaCrawler. Each of them has a unique interface and different databases covering various portions of the WWWeb. We chose to use the MetaCrawler Selberg and Etzioni 1995 for the Internet search, due to its parallel search capabilities, that can be ported to a future multithreaded Java implementation that we envision. Such an approach will ideally use Java applets for the multithreaded propagation rules of PARIS Harabagiu and Moldovan 1996*d*, our parallel inference system.

The syntax of the MetaCrawler interface permits: (a) a search for words as a phrase, (b) a search for a list of words with an equivalent logical 'and' between the words or (c) a search for a list of words with an equivalent logical 'or' between the words. The contextual broker has to translate the context into a source model presented to the MetaCrawler entries. The algorithm that provides with lists of words and phrases presented to the crawler is similar to the Internet Learning Agent presented in Perkowitz and Etzioni (1995).

The queries posted by the Metacrawler are accessed with shallow processing techniques, to retrieve text paragraphs that are assumed relevant. When the response text is available, its local and global contexts are determined by the inference system.

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

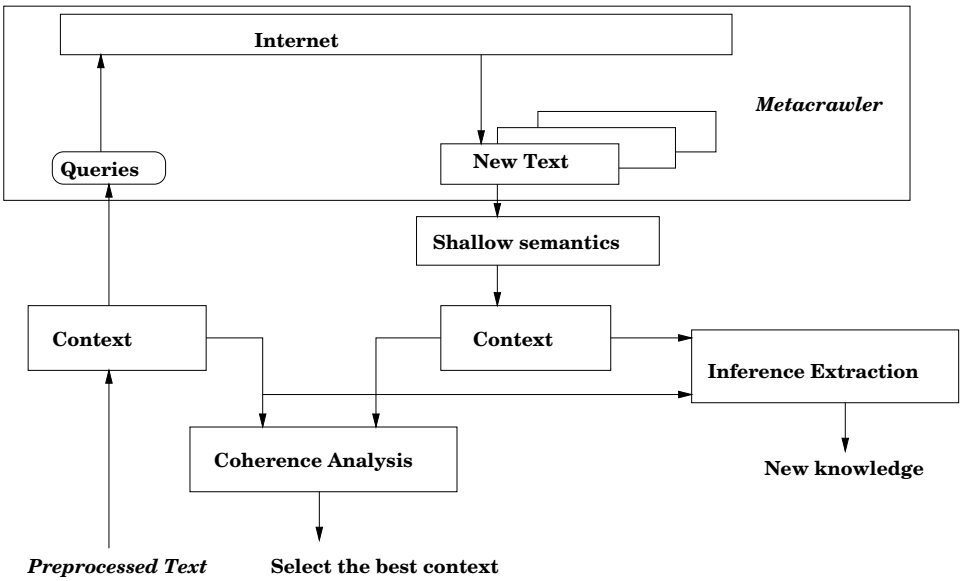


Fig. 5. The mechanism for acquiring knowledge from Internet.

Then, its relevance is evaluated using a coherence measure that adapts the metric from Ng and Mooney 1990 to the multibranch path representation of context.

Figure 5 shows the mechanism for acquiring information from Internet related to the initial text. The process starts with the extraction of context from the initial text as it was discussed in the previous section. Then, queries are formed with the concepts of interest. The MetaCrawler submits these queries to the Internet and as a result several text documents are extracted. The TextNet system then extracts the context from each of the new texts that is related to the context of the original text. The coherence analysis module performs an evaluation, as it will be described below, and selects the most appropriate new context.

The relevance of the new texts fetched by the MetaCrawler depends in great part of the queries presented to the Internet. Two problems have been identified in formulating these queries. One is the *discovery* problem, i.e what are the queries that best lead to the discovery of additional world knowledge. The other is the *quality* problem, i.e. what is the relevance of the information to the context under consideration. Both these problems can be traced to the inherent difficulty of lexical ambiguity of natural language. Sometimes we may use directly the concepts that appear in the text and other times we may use other concepts that are semantically close to them. For example, given the context between sentences `S1` and `S2` we expect to obtain additional knowledge that motivates the relation between the verb `hit#1-14:v` and the noun `baseline#1-1:n`. For this we may form a query that contains the word `tennis`, since it is likely that will bring new information relevant to tennis.

Conversely, once that a text is retrieved, the ambiguous words are disambiguated

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

Table 4. *Some of MetaCrawler returns*

http://www.xmission.com/~gastown/tennis/thonda4.html
http://www.netvideo.com/mediamart/video/tennis/sku/mm102744.html
http://sports.yahoo.com/ten/lgns/961205/womenswin.html
http://jtta.ge.niigata-u.ac.jp

using shallow semantics, based on the assumption that the words used in the query were already disambiguated, and the new words belong to synsets semantically close to those from the context. The recognition of the lexical relations uses a set of heuristics developed for the transformation of WordNet glosses.

5.2 Coherence evaluation

Although the new text is assumed to be coherent in itself , there is no guarantee that the information it brings is coherent with the initial context structure. We propose a method to evaluate the relevance of the new information. For this we define a coherence metric as follows:

$$\mathscr{C} = \frac{N^{common} \cdot L^{common}}{N^{total} \cdot L^{total}}$$

where N^{common} is the number of common concept nodes between the novel context and the original one. Similarly L^{common} represents the length of the common semantic connections shared by both contextual structures. N^{total} is the total number of nodes in both contexts and L^{total} counts all the links spanning both contexts.

This measure is constructed so that it increases with the number of common concepts and semantic connections. The denominator scales the disjunctive part of the two concepts with the size of the total contexts.

To illustrate the computation of the coherence metric, let us consider the following to be the initial text:

```
S3: The kid was playing too much.
S4: He hit the ball very hard.
S5: It landed almost always near the baseline.
```

The query presented to the MetaCrawler is [`play + tennis * ball * baseline + hit`]. Table 4 illustrates some of the URLs of the crawler's return. The coherence measure identified the closest context as being retrieved from a text that describes tennis winning strategies. The closest paragraph from the technical report entitled "*Doubles Unorthodox Offensive Tactics and Strategies*", by Farrell Honda, obtainable from the URL:`http://www.xmission.com/~`*gastown/tennis/thonda4.html* is:

Successful unorthodox serves near the baseline will result in one of two possibilities. The receiver will be aced; the receiver will be stunned into losing a fraction of second in reaction time and hit the ball into the net.

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

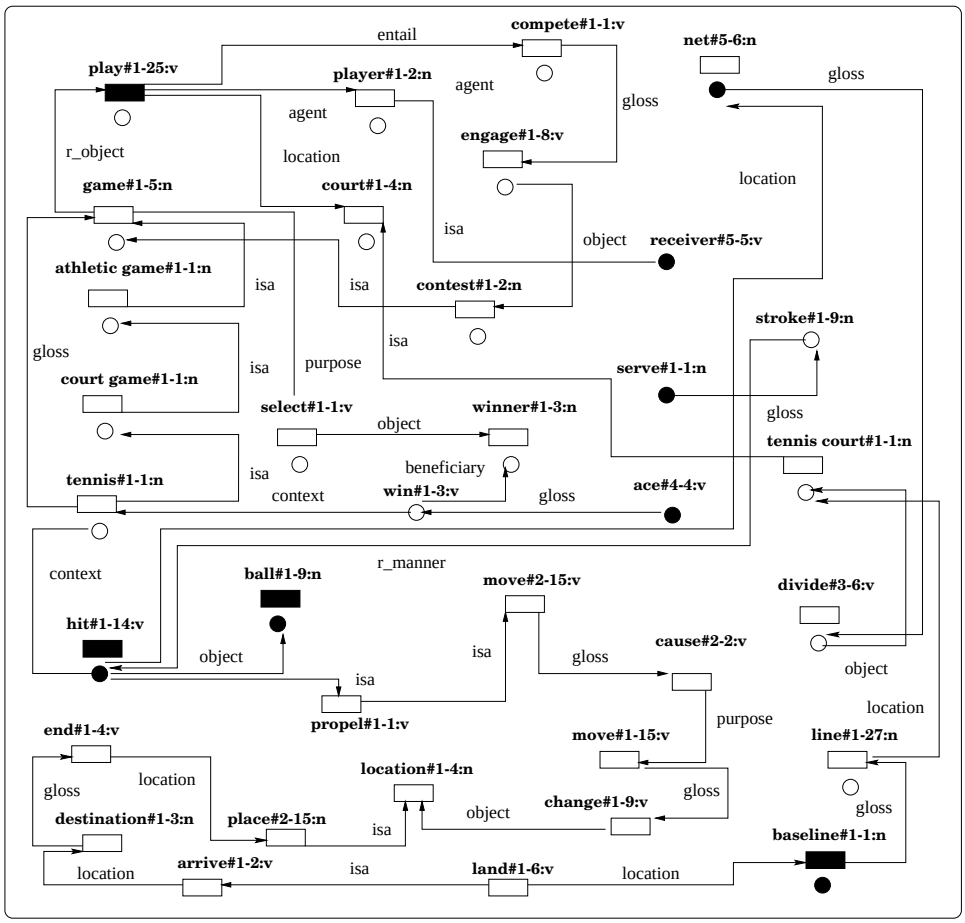


Fig. 6. Two complementary contextual structures.

Figure 6 illustrates the two contextual structures overlapped. The concepts marked with a rectangle belong to the initial text context and those marked with a circle belong to the context derived from the Internet text. The common concepts have both a rectangle and a circle. The concepts that are blackened appear in texts, the others are only knowledge base concepts.

As we can see from the figure $N^{common} = 18$, $L^{common} = 21$, $N^{total} = 34$ and $L^{total} = 38$. It results a coherence factor $\mathscr{C} = 18 \text{ x } 21/\ 34 \text{ x } 38 = 0.29$. The two contexts have common textual concepts `ball`, `hit` and `baseline`. Therefore all connections between these concepts are common to both contexts. Moreover, two of the other textual concepts (`receiver` and `net`) of the novel context belong to the first context (or are hyponyms of concepts from the first context). The remaining textual concepts of the novel context (`ace` and `serve`) deliver semantic connections that involve concepts and links that were already present in the first context.

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

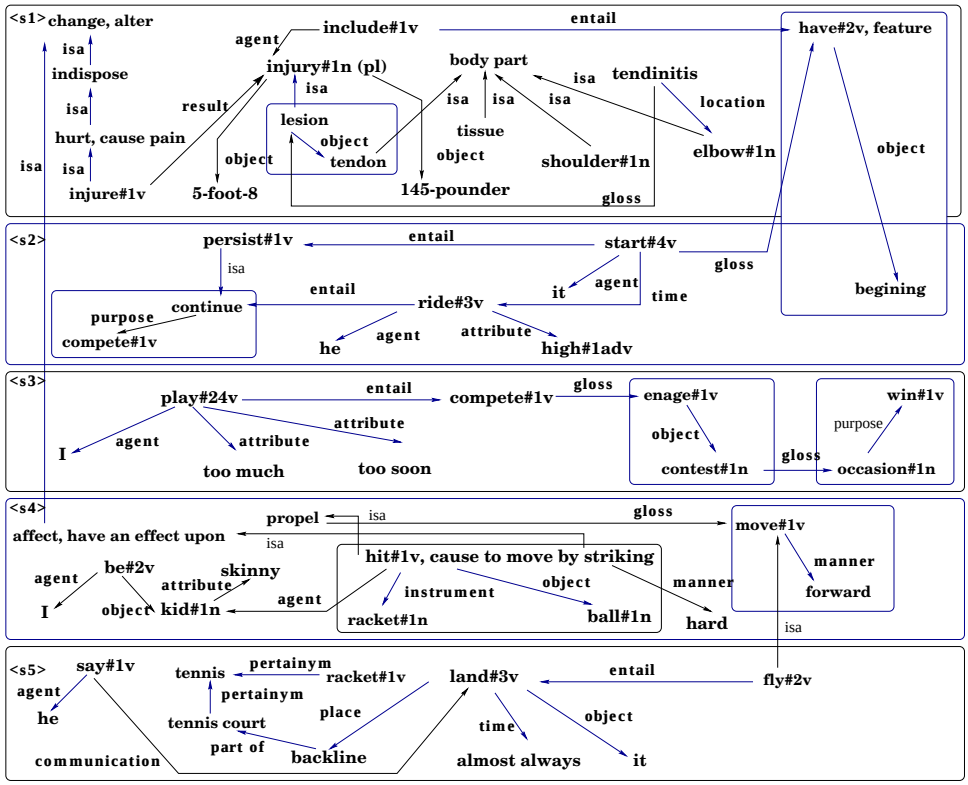


Fig. 7. Semantic paths in a real world text.

6 An example

In this section we consider a real text as an input and show how additional knowledge may be acquired from Internet that supplements the one that can be inferred from a knowledge base such as extended WordNet. We shall illustrate this claim by recognizing the compound concept `injuries related to sports`. We present a typical text whose topic is `injuries related to sports`:

```
<s1> The 5-foot-8, 145-pounder's injuries included tendinitis in his
    shoulder and elbow.
<s2> It started when he was riding high, and has persisted.
<s3> "I guess I played too much too soon.
<s4> I was a skinny kid who hit the ball very hard.
<s5> Maybe too hard, it landed almost always near the baseline" he says.
```

Figure 7 illustrates some of semantic paths found by the inference system for this text. This is a paragraph from a *Treebank Wall Street Journal* article Marcus *et al.* 1993 reporting tennis-related events at the U.S. Open. The text provides a fine non-trivial pragmatic inference example. The blank squares in Figure 7 represent the glosses connected to their corresponding concepts through *gloss* relations.

Sentences `<s3>`, `<s4>` and `<s5>` have provided inspiration for the text `S3--S5`

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

presented in the previous section, and the reader is already familiar to their context. The global context also contains explanatory relations between the injuries of the player and his string of competitions and intention of winning.

The inference system managed to recognize fairly well the global context which supports the informational aspects of this discourse. As the current knowledge base lacks a representation for plans and the intentions they underlie, the inference scheme failed to recognize subgoals or intensional information.

Contextual brokering is required to provide with information that motivates the semantic bridging of the five sentences. Multiple queries involving concepts from the five sentences of the text were submitted to the crawler. The most coherent context that was returned connects the verbs `win` and `hit`. The identification of the goal of hitting is thus determined in the intension of wining the tennis match.

The contextual pointers from the glosses of `ace#4n` and `tennis`, extended with the additional information from the local context of the Internet text, added to the entailment between `play` and `win` account for the recognition of the goal of the text. Winning the game is the motivation of the actions determining injuries in tennis (the topic of the text). Altogether, TextNet creates on the fly a frame for the compound concept `injuries related to sports` by using the results of the contextual brokering. The frame comprises connections to the other forms of additional knowledge (e.g. the plans and the goals):

---

injuries related to tennis are caused by:
- ○ *playing too much tennis (too soon)*
- ○ *wanting to win the match (GOAL)*
  - → *PLAN: hit the ball hard*
  - —*entails*→ *the ball lands near the baseline*
    - —*entails*→ *the competitor can't reach the ball*
      - —*entails*→ *the player wins the point (SUBGOAL)*

forms of injuries related to tennis:
- ○ *tendinitis in the shoulder and elbow*

player's condition:
- ○ *skinny (145 pounds, 5-foot-8); young (kid)*

The previous text is also linked to the following two messages, describing injuries in cowboy polo and in gymnastics. The system infers conditions in which injuries may occur in sports, and connects the compound concept to the plans and the goals that were followed when the injuries occurred. Other contextual information is also available, like the physical condition of the player, or special events that may cause typical injuries, e.g. falling from the horizontal bar may lead to a broken leg.

```
<s> Since he began playing cowboy polo 26 years ago, Eddie Sorensen has
had four operations, a broken leg, several cracked ribs, a separated
shoulder, and a batch of concussions.
<s> Once he got whacked under the jaw with a mallet, leaving a perfect
imprint of all his teeth on his tongue.
```

```
<s> Tim Daggett, a holdover from '84, broke a leg so badly last
year that amputation was considered.
<s> Dan Hayden took over the No. 1 rating during Daggett's absence, and won
```

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

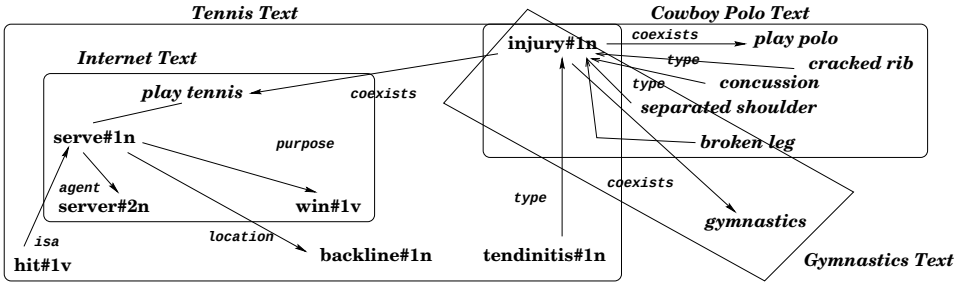


Fig. 8. Contextual connections between texts.

```
the U.S. National meet last month that counted 40% toward Olympic
qualification.
<s> Shockingly, however, he fell twice on his last event Friday, the
horizontal bar, and his score of 8.3 plummeted him from second to eighth.
<s> He later revealed that he'd separated a shoulder in a previous exercise.
```

Figure 8 illustrates some contextual concepts that connect the texts in an informational web. The tennis text indexes the Internet text, using coreference of concept relations: e.g. the *location* case function between `serve` (or `hit`) and `baseline`, occurring in both texts. Furthermore, the tennis text is connected to the polo text and to the gymnastics text, as they all have in common injuries occurring while practicing sports.

7 Discussion

There is a growing trend to use softbots to get information from the web. Several sophisticated Internet interfaces have emerged, including SIMS Knoblock *et al.* 1994, the Information Manifold at AT&T Kirk *et al.* 1995 and the Internet softbot Etzioni and Weld 1994. The use of AI techniques and of natural language methods for this purpose is yet in an incipient stage. In this paper, we have argued that the Internet provides a powerful platform from which common-sense knowledge can be gathered. Our client is TextNet, a system that combines the principled simplicity of WordNet with a web of contextually indexed texts. Many interesting aspects of knowledge acquisition, like the treatment of inconsistencies and information distortions have yet to be addressed.

There is a growing trend to use electronic dictionaries to construct large linguistic knowledge bases. In this paper we argued that WordNet provides a platform on which powerful, generally applicable knowledge bases can be built. Although WordNet does not provide information regarding plans and the supporting goals, or conditions necessary for actions and events that may be necessary to fully explain a text, it nevertheless contains almost all English words mapped into concepts and a large number of links between concepts. Extended with micro-contextual defining features, we found WordNet to be an ideal setting on which to build a representation for complex, structured concepts.

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

References

Ali, S. and Shapiro, S. C. (1993) Natural Language Processing Using Propositional Semantic Network with Structured Variables. *International Journal of Minds and Machines* (Special Issue on 'Knowledge Representation for Natural Language'), **3**(4): 421–453.

Broglio, J., Callan, J. and Croft, W. B. (1993) The INQUIRY System. In: R. Merchant (ed.), *Proceedings of the TIPSTER Text Program – Phase I*. Morgan Kaufmann.

Etzioni, O. and Weld, D. (1994) A softbot-based interface to Internet. *Communications of the ACM* **37**(7): 72–76.

Etzioni, O. (1996) Moving up the information food chain: deploying softbots on the World Wide Web. *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, pp. 1322–1326.

Fellbaum, C., Gross, D. and Miller, G. A. (1991) A lexical database organized on psycholinguistic principles. In: U. Zernik (ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pp. 141–170. Lawrence Erlbaum.

Fikes, R., Farquhar, A. and Pratt, W. (1996) Information Brokers: Gathering Information from Heterogenous Information Sources. *Proceedings of the Ninth Florida Artificial Intelligence Research Symposium*, pp. 192–197.

Harabagiu, S. and Moldovan, D. (1995) A Parallel Algorithm for Text Inference. *Working Notes of the Workshop on Parallel Processing for Artificial Intelligence: IJCAI-95*, Montreal, Canada, pp. 76–86.

Harabagiu, S. and Moldovan, D. (1996*a*) A Parallel Algorithm for Text Inference. *Proceeding of the International Parallel Processing Symposium*, pp. 441–445.

Harabagiu, S. and Moldovan, D. (1996*b*) Testing Gricean Constraints on a WordNet-based Coherence Evaluation System. *Working Notes of the AAAI-96 Spring Symposium on Computational Implicature*, Stanford, CA, pp. 31–38.

Harabagiu, S. and Moldovan, D. (1996*c*) TextNet – a text-based intelligent system. *Working Notes of the AAAI-96 Fall Symposium on Knowledge Representation Systems based on Natural Language*, MIT, Cambridge, MA, pp. 32–43.

Harabagiu, S. and Moldovan, D. (1996*d*) A Parallel Inference System. *Proceedings of the 8th International Conference on Tools with Artificial Intelligence*, Toulouse, France, pp. 216–223.

Harabagiu, S. and Moldovan, D. (1997*a*) Knowledge Processing on Extended WordNet. In: C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database and Some of its Applications*. MIT Press.

Harabagiu, S. and Moldovan, D. (1997*b*) A Parallel Inference System. *IEEE Transactions on Parallel and Distributed Systems*. Submitted.

Harabagiu, S. (1996) An Application of WordNet to Prepositional Attachment. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, CA, pp. 360–362.

Iwanska, L. (1992) A General Semantic Model of Negation in Natural Language: Representation and Inference. *Proceedings of Third International Conference on Principles of Knowledge Representation and Reasoning (KR-92)*, pp. 357–368.

Jacobs, P. S. (1992) Text Power and Intelligent Systems. In: P. S. Jabobs (ed.), *Text–Based Intelligent Systems*, pp. 1–8. Lawrence Erlbaum.

Kirk, T., Levy, A. Y., Sagiv, Y. and Srivastava, D. (1995) The information manifold. *Working Notes of the AAAI-95 Spring Symposium on Information Gathering from Heterogenous Distributed Environments*, Stanford, CA, pp. 85–91.

Knoblock, C., Arens, Y. and Hsu, C. N. (1994) Cooperating agents for information retrieval. *Proceedings of Second International conference on Cooperative Information systems*, Toronto, Canada.

Lenat, D. B. (1995) CYC: A Large-Scale Investment in Knowledge Infrastructure. *Communication of the ACM*, **38**(11): 32–39.

Liddy, E. D. and Myaeng, S. H. (1993) A system update for TREC-2. *Proceedings of the TIPSTER Text Program – Phase I*, San Mateo, CA.

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press

Lin, C. E., Harabagiu, S. M., Hendrickson, K. J. and Moldovan, D. I. (1994) Marker Propagation Networks and Their Implementation. *Technical Report*, Southern Methodist University.

Marcus, M., Santorini, B. and Marcinkiewicz, M. A. (1993) Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, **19**(2): 313–330.

Moldovan, D. I., Lee, W. and Lin, C. E. (1992) SNAP: A Marker-Propagation Architecture for Knowledge Processing. *IEEE Transactions on Parallel and Distributed Systems*, **3**: 1–13.

Miller, G. A. (1995) WordNet: A Lexical Database. *Communication of the ACM*, **38**(11): 39–41.

Ng, H. T. and Mooney, R. J. (1990) On the Role of coherence in abductive explanation. *Proceedings of the Eigth National Conference on Artificial Intelligence (AAAI-90)*, pp. 337–342.

Perkowitz, M. and Etzioni, O. (1995) Category Translation: learning to understand information on the Internet. *Proceedings of the International Joint Conference on Artificial Intelligence, (IJCAI-95)*, pp. 930–936.

Selman, B., Brooks, R. A., Dean, R., Horvitz, E., Mitchell, T. M. and Nilsson, N. J. (1996) Challenge problems for Artificial Intelligence. *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, pp. 1340–1346.

Selberg, E. and Etzioni, O. (1995) Multi-service search and comparison using the MetaCrawler. *Proceedings of the 4th World Wide Web Conference*, pp. 195–208.

Shakes, J., Langheinrich, M. and Etzioni, O. (1996) Ahoy! the home page finder. *Technical Report*, University of Washington.

Schubert, L. K. and Hwang, C. H. (1989) An Episodic Knowledge Representation for Narrative Texts. *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning (KR-89)*, pp. 444–458.

Sowa, J. F. (1984) *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley.

Sowa, J. F. (1991) Toward the Expressive Power of Natural Language. In: J. F. Sowa (ed.), *Principles of Semantic Networks*, pp. 157–190. Morgan Kaufmann.

Sundheim, B. (1993) *Proceedings of the Fifth Message Understanding Conference (MUC-5)*.

https://doi.org/10.1017/S1351324997001757 Published online by Cambridge University Press