

Hybrid Systems for Information Extraction and Question Answering

Answering

Rodolfo Delmonte

Ca' Bembo, San Trovaso 1075

Università "Ca Foscari"

30123 - VENEZIA

Tel. 39-041-2345717/12 - Fax. 39-041-2345703

E-mail: delmont@unive.it - website: project.cgm.unive.it

Abstract

Information Extraction, Summarization and Question Answering all manipulate natural language texts and should benefit from the use of NLP techniques. Statistical techniques have till now outperformed symbolic processing of unrestricted text. However, Information Extraction and Question Answering require by far more accurate results of what is currently produced by Bag-Of-Words approaches. Besides, we see that such tasks as Semantic Evaluation of Text Entailment or Similarity – as required by the RTE Challenge, impose a much stricter performance in semantic terms to tell true from false pairs. We will speak in favour of a hybrid system, a combination of statistical and symbolic processing with reference to a specific problem, that of Anaphora Resolution which looms large and deep in text processing.

1. Introduction

Although full syntactic and semantic analysis of open-domain natural language text is beyond current technology, a number of papers have been recently published [1,2,3] showing that, by using probabilistic or symbolic methods, it is possible to obtain dependency-based representations of unlimited texts with good recall and precision. Consequently, we believe it should be possible to augment the manual-annotation-based approach with automatically built annotations by extracting a limited subset of semantic relations from unstructured text. In short, shallow/partial text understanding on the level of semantic relations, an extended label including Predicate-Argument Structures and other syntactically and semantically derivable head modifiers and adjuncts. This approach is promising because it attempts to address the well-known shortcomings of standard “bag-of-words” (BOWs) information retrieval/extraction techniques without requiring manual intervention: it develops current NLP technologies which make heavy use of statistically and FSA based approaches to syntactic parsing.

GETARUNS [4,5,6], a text understanding system (TUS), developed in collaboration between the University of Venice and the University of Parma, can perform semantic analysis on the basis of syntactic parsing and, after performing anaphora resolution, builds a quasi

logical form with flat indexed Augmented Dependency Structures (ADSs). In addition, it uses a centering algorithm to individuate the topics or discourse centers which are weighted on the basis of a relevance score. This logical form can then be used to individuate the best sentence candidates to answer queries or provide appropriate information.

This paper is organized as follows: in section 2 below we discuss why deep linguistic processing is needed in Information Retrieval and Information Extraction; in section 3 we present GETARUNS, the NLP system and the Upper Module of GETARUNS; in section 4 we describe two experiments with state-of-the-art benchmark corpora.

2 Ternary Expressions as Predicate-Argument Structures

Researchers like Lin, Katz and Litkowski have started to work in the direction of using NLP to populate a database of RDFs, thus creating the premises for the automatic creation of ontologies to be used in the IR/IE tasks. However, in no way RDFs and ternary expressions may constitute a formal tool sufficient to express the complexity of natural language texts.

RDFs are assertions about the things (people, Webpages and whatever) they predicate about by asserting that they have certain properties with certain values. If we may agree with the fact that this is natural way of dealing with data handled by computers most frequently, it also a fact that this is not equivalent as being useful for natural language. The misconception seems to be deeply embedded in the nature of RDFs as a whole: they are directly comparable to attribute-value pairs and DAGs which are also the formalism used by most recent linguistic unification-based grammars. From the logical and semantic point of view RDFs also resemble very closely first order predicate logic constructs: but we must remember that FOPL is as such insufficient to describe natural language texts.

Ternary expressions(T-expressions), <subject relation object>.

Certain other parameters (adjectives, possessive nouns, prepositional phrases, etc.) are used to create additional T-expressions in which prepositions and several special words may serve as relations. For instance, the following simple sentence

(1) Bill surprised Hillary with his answer

will produce two T-expressions:

(2) <<Bill surprise Hillary> with answer>
<answer related-to Bill>

In Litkowski's system the key step in their question-answering prototype was the analysis of the parse trees to extract semantic relation triples and populate the databases used to answer the question. A semantic relation triple consists of a discourse entity, a semantic relation which characterizes the entity's role in the sentence, and a governing word to which the entity stands in the semantic relation. The semantic relations in which entities participate are intended to capture the semantic roles of the entities, as generally understood in linguistics. This includes such roles as agent, theme, location, manner, modifier, purpose, and time. Surrogate place holders included are "SUBJ," "OBJ," "TIME," "NUM," "ADJMOD," and the prepositions heading prepositional phrases. The governing word was generally the word in the sentence that the discourse entity stood in relation to. For "SUBJ," "OBJ," and "TIME," this was generally the main verb of the sentence. For prepositions, the governing word was generally the noun or verb that the prepositional phrase modified. For the adjectives and numbers, the governing word was generally the noun that was modified.

2.1 Ternary Expressions are better than the BOWs approach, but...

People working advocating the supremacy of the Tes approach were reacting against the Bag of Words approach of IR/IE in which words were wrongly regarded to be entertaining a meaningful relation simply on the basis of topological criteria: normally the distance criteria or the more or less proximity between the words to be related. Intervening words might have already been discarded from the input text on the basis of stopword filtering. Stopwords list include all grammatical close type words of the language considered useless for the main purpose of IR/IE practitioners seen that they cannot be used to denote concepts. Stopwords constitute what is usually regarded the noisy part of the channel in information theory. However, it is just because the redundancy of the information channel is guaranteed by the presence of grammatical words that the message gets appropriately computed by the subject of the communication process, i.e. human beings. Besides, entropy is not to be computed in terms of number of words or letters of the alphabet, but in number of semantic and syntactic relation entertained by open class words (nouns, verbs, adjectives, adverbials) basically by virtue of closed class words. Redundancy should then be computed on the basis of the ambiguity intervening when enumerating those relations, a very hard task to accomplish which has never been attempted yet, at least to my knowledge.

What people working with TEs noted was just the problem of encoding relations appropriately, at least some of these relations. The IR/IE BOWs approach

suffers (at least) from Reversible Arguments Problem (see [7])

- What do frogs eat? vs What eats frogs?

The verb "eat" entertains asymmetrical relations with its SUBJECT and its OBJECT: in one case we talk of the "eater", the SUBJECT and in another case of the "eatee", the OBJECT. Other similar problems occur with TEs when the two elements of the relation have the same head, as in:

-The president of Russia visited the president of China.
Who visited the president?

The question will not be properly answered in lack of some clarification dialogue intervening, but the corresponding TEs should have more structure to be able to represent the internal relations of the two presidents. The asymmetry of relation in transitive constructions involving verbs of accomplishments and achievements (or simply world-changing events) is however further complicated by a number of structural problems which are typically found in most languages of the world, the first one and most common being Passive constructions:

i. John killed Tom.

ii. Tom was killed by a man.

Who killed the man?

Answer to the question would be answered by "John" in case the information available was represented by sentence in i., but it would be answered by "Tom" in case the information available was represented by sentence ii. Obviously this would happen only in lack of sufficient NLP elaboration: a too shallow approach would not be able to capture presence of a passive structure. We are here referring to "Chunk"-based approaches those in which the object of computation is constituted by the creation of Noun Phrases and no attempt is made to compute clause-level structure.

There is a certain number of other similar structure in texts which must be regarded as inducing into the same type of miscomputation: i.e. taking the surface order of NPs as indicating the deep intended meaning. In all of the following constructions the surface subject is on the contrary the deep object thus the Affected Theme or argument that suffers the effects of the action expressed by the governing verb rather than the Agent:

Inchoatized structures; Ergativized structures; Impersonal structures

Other important and typical structures which constitute problematic cases for a surface chunks based TEs approach to text computation are the following ones in which one of the arguments is missing and Control should be applied by a governing NP, they are called in one definition Open Predicative structures and they are

Relative clauses; Fronted Adjectival adjunct clauses; Infinitive clauses; Fronted Participial clauses;; Gerundive Clauses; Elliptical Clauses; Coordinate constructions

In addition to that there is one further problem and is definable as the Factuality Prejudice: by collecting

keywords and TEs people apply a Factuality Presupposition to the text they are mining: they believe that all terms being recovered by the search represent real facts. This is however not true and the problem is related to the possibility to detect in texts the presence of such semantic indicators as those listed here below:

Negation; Quantification; Opaque contexts (wish, want); Future, Subjunctive Mode; Modality; Conditionals

Finally there is a discourse related problem and is the **Anaphora Resolution** problem which is the hardest to be tackled by NLP: it is a fact that anaphoric relations are the building blocks of cohesiveness and coherence in texts. Whenever an anaphoric link is missed one relation will be assigned to a wrong referring expression thus presumably jeopardising the possibility to answer a related question appropriately. This is we believe the most relevant topic to be put forward in favour of the need to have symbolic computational linguistic processing (besides statistical processing).

3 GETARUNS – the NLUS

GETARUN, the System for Natural Language Understanding, produces a semantic representation in xml format, in which each sentence of the input text is divided up into predicate-argument structures where arguments and adjuncts are related to their appropriate head. Consider now a simple sentence like the following:

(1) John went into a restaurant

GETARUNS represents this sentence in different manners according to whether it is operating in Complete or in Shallow modality. In turn the operating modality is determined by its ability to compute the current text: in case of failure the system will switch automatically from Complete to Partial/Shallow modality.

The system will produce a representation inspired by Situation Semantics[14] where reality is represented in Situations which are collections of Facts: in turn facts are made up of Infons which are information units characterised as follows:

*Infon(Index,
Relation(Property),
List of Arguments - with Semantic Roles,
Polarity - 1 affirmative, 0 negation,
Temporal Location Index,
Spatial Location Index)*

In addition each Argument has a semantic identifier which is unique in the Discourse Model and is used to individuate the entity uniquely. Also propositional facts have semantic identifiers assigned, thus constituting second level ontological objects. They may be “quantified” over by temporal representations but also by discourse level operators, like subordinating conjunctions and a performative operator if needed. Negation on the contrary is expressed in each fact.

In case of failure at the Complete level, the system will switch to Partial and the representation will be deprived of its temporal and spatial location information. In the current version of the system, we use Complete modality

for tasks which involve short texts (like the students summaries and text understanding queries), where text analyses may be supervised and updates to the grammar and/or the lexicon may be needed. For unlimited text from the web we only use partial modality. Evaluation of the two modalities are reported in a section below.

3.1 The Parser and the Discourse Model

As said above, the query building process needs an ontology which is created from the translation of the Discourse Model built by GETARUNS in its Complete/Partial Representation. GETARUNS, is equipped with three main modules: a lower module for parsing where sentence strategies are implemented; a middle module for semantic interpretation and discourse model construction which is cast into Situation Semantics; and a higher module where reasoning and generation takes place. The system works in Italian and English.

Our parser is a rule-based deterministic parser in the sense that it uses a lookahead and a Well-Formed Substring Table to reduce backtracking. It also implements Finite State Automata in the task of tag disambiguation, and produces multiwords whenever lexical information allows it. In our parser we use a number of parsing strategies and graceful recovery procedures which follow a strictly parameterized approach to their definition and implementation. A shallow or partial parser is also implemented and always activated before the complete parse takes place, in order to produce the default baseline output to be used by further computation in case of total failure. In that case partial semantic mapping will take place where no Logical Form is being built and only referring expressions are asserted in the Discourse Model – but see below.

3.2 Lexical Information

The output of grammatical modules is then fed onto the Binding Module(BM) which activates an algorithm for anaphoric binding in LFG (see [13]) terms using f-structures as domains and grammatical functions as entry points into the structure. We show here below the architecture of the system. The grammar is equipped with a lexicon containing a list of 30000 wordforms derived from Penn Treebank.

However, morphological analysis for English has also been implemented and used for OOV words. The system uses a core fully specified lexicon, which contains approximately 10,000 most frequent entries of English. In addition to that, there are all lexical forms provided by a fully revised version of COMLEX. In order to take into account phrasal and adverbial verbal compound forms, we also use lexical entries made available by UPenn and TAG encoding. Their grammatical verbal syntactic codes have then been adapted to our formalism and is used to generate an approximate subcategorization scheme with an approximate aspectual class associated to it.

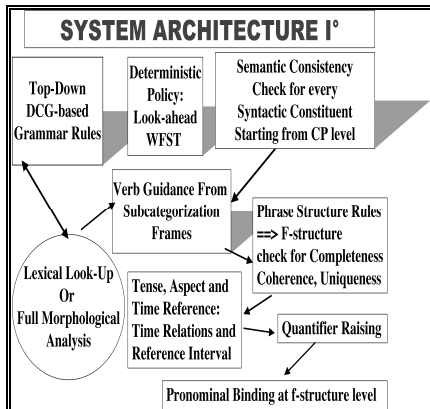


Fig. 1. GETARUNS' LFG-Based Parser

Semantic inherent features for Out of Vocabulary words, be they nouns, verbs, adjectives or adverbs, are provided by a fully revised version of WordNet – 270,000 lexical entries - in which we used 75 semantic classes similar to those provided by CoreLex. Subcategorization information and Semantic Roles are then derived from a carefully adapted version of FrameNet and VerbNet. Our “training” corpus is made up of 200,000 words and contains a number of texts taken from different genres, portions of the UPenn Treebank corpus, test-suits for grammatical relations, and sentences taken from COMLEX manual. An evaluation carried out on the Susan Corpus related GREVAL test suite made of 500 sentences has been reported lately [12] to have achieved 90% F-measure over all major grammatical relations. We achieved a similar result with the shallow cascaded parser, limited though to only SUBJECT and OBJECT relations on LFG-XEROX 700 corpus.

3.3 The Upper Module

GETARUNS, as shown in Fig.2 has a linguistically-based semantic module which is used to build up the Discourse Model. Semantic processing is strongly modularized and distributed amongst a number of different submodules which take care of Spatio-Temporal Reasoning, Discourse Level Anaphora Resolution, and other subsidiary processes like Topic Hierarchy which will impinge on Relevance Scoring when creating semantic individuals. These are then asserted in the Discourse Model (hence the DM), which is then used to solve nominal coreference together with WordNet. Semantic Mapping is performed in two steps: at first a Logical Form is produced which is a structural mapping from DAGs onto of unscoped well-formed formulas. These are then turned into situational semantics informational units, infons which may become facts or sits.

In each infon, Arguments have each a semantic identifier which is unique in the DM and is used to individuate the entity. Also propositional facts have semantic identifiers assigned thus constituting second level ontological objects. They may be “quantified” over by temporal representations but also by discourse level operators, like subordinating conjunctions. Negation on the contrary is

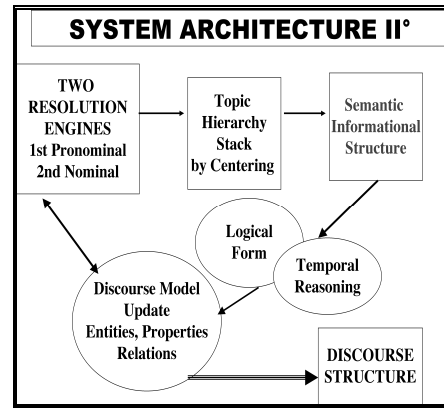


Fig. 2. GETARUNS' Discourse Level Modules

expressed in each fact. All entities and their properties are asserted in the DM with the relations in which they are involved; in turn the relations may have modifiers - sentence level adjuncts and entities may also have modifiers or attributes. Each entity has a polarity and a couple of spatiotemporal indices which are linked to main temporal and spatial locations if any exists; else they are linked to presumed time reference derived from tense and aspect computation. Entities are mapped into semantic individuals with the following ontology: on first occurrence of a referring expression it is asserted as an INDIVIDUAL if it is a definite or indefinite expression; it is asserted as a CLASS if it is quantified (depending on quantifier type) or has no determiner. Special individuals are ENTs which are associated to discourse level anaphora which bind relations and their arguments. Finally, we have LOCs for main locations, both spatial and temporal. Whenever there is cardinality determined by a digit, its number is plural or it is quantified (depending on quantifier type) the referring expression is asserted as a SET. Cardinality is simply inferred in case of naked plural: in case of collective nominal expression it is set to 100, otherwise to 5. On second occurrence of the same nominal head the semantic index is recovered from the history list and the system checks whether it is the same referring expression:

- in case it is definite or indefinite with a predicative role and no attributes nor modifiers, nothing is done;
- in case it has different number - singular and the one present in the DM is a set or a class, nothing happens;
- in case it has attributes and modifiers which are different and the one present in the DM has none, nothing happens;
- in case it is quantified expression and has no cardinality, and the one present in the DM is a set or a class, again nothing happens.

In all other cases a new entity is asserted in the DM which however is also computed as being included in (a superset of) or by (a subset of) the previous entity.

The upper module of GETARUNS has been evaluated on the basis of its ability to perform anaphora resolution and to individuate referring expressions, with a corpus of 40,000 words: it achieved 74% F-measure.

4. Two experiments with GETURANS

As an example of the shallow system we discuss here below the analysis of a newspaper article which as would usually be the case has a certain number of pronominal expressions, which modify the relevance of lexical descriptions in the overall processing for the search of either "Named Entities" or simply entities individuated by common nouns. If the count is based solely on lexical lemmata and not on the presence of coreferential pronominal expressions, the results will be heavily biased and certainly wrong. Here is the text:

1. Thursday, 25th June 2001
National Parties and the Internet
by Joanna Crawford
2. A survey of how national parties used the internet as a campaigning tool during the election will brand **their** efforts "bleak and dispiriting" - despite the pre-campaign hype of an "e-election".
3. Researchers from Salford University studied websites from all the major parties during the general election, as well as looking at every site put up by local candidates.
4. **Their** conclusions - to be presented tomorrow at a special conference organised by the Institute for Public Policy Research - could influence how future political contests, including the forthcoming Euro debate, are carried out on the web.
5. The report finds that **none** of the major three parties allowed message boards or chat rooms for users to post their opinions on the sites.
6. **It** states: "Parties were accused of simply engaging in online propaganda with boring content and largely ignoring interactivity."
7. The report concludes: "The new media is a way for **them** to get closer to the public without necessarily allowing the public to become overly familiar in return.
8. The authors - Rachel Gibson and Stephen Ward - go on to state that **this** may be because parties still regard the web as an electioneering tool, rather than as a democratic device.
9. **They** said: "Very **few** offered original material, or changed **their** sites noticeably over the course of the campaign.
10. Indeed, a large **majority** of local sites were really no more than static electronic brochures."
11. **They** dub **this** "rather disappointing", but praise the Liberal Democrats as "clearly the most active" with around 150 sites. The report concludes: "Parties, as with the general public, need incentives to use the technology.
12. As yet, there seems more to lose and less to gain if **they** make mistakes experimenting with the technology."

We highlighted pronominal expressions in bold. In a BOWs approach, the count for most relevant topics is solely based on lexical descriptions and "party, internet" are computed as the most important key-words. However, after the text has been passed by the partial semantic analysis, "researcher, author" come up as important topics.

We report here below the output of the Anaphora Resolution module: in interaction with the Discourse Model where semantic indices are asserted for each entity. Sentence numbers are taken from the text. We report Anaphora Resolution decisions: in particular in sentences where a

pronoun is coreferred to an antecedent, the antecedent is set as current Main Topic and its semantic ID is used.

1. state(1, change)
topics: main:party, secondary: internet
topics(1, main, id1; secondary, id2; potential, id3)
2. state(2, continue)
topics: main:party, secondary: survey
topics(2, main, id1; secondary, id7; potential, id2)
3. state(3, retaining)
topics: main: researcher, secondary: party
topic(3, main, id18; secondary, id1; , id19)
4. Anaphora Resolution: their resolved as researcher
state(4, continue)
topics: main: researcher, secondary: contest
topics(4, main, id18; secondary, id26; potential, id27)
5. state(5, retaining)
topics: main: report, secondary: researcher
topics(5, main, id7; secondary, id18; potential, id1)
6. Anaphora Resolution: it resolved as report
state(6, continue)
topics: main: report, secondary: party
topics(6, main, id7; secondary, id1; potential, id40)
7. state(7, continue)
topics: main: report, secondary: party
topics(7, main, id7; secondary, id1; potential, id2)
8. The authors - Rachel Gibson and Stephen Ward - go on to state that this may be because parties still regard the web as an electioneering tool, rather than as a democratic device.
Anaphora Resolution: this resolved as 'discourse bound'
state(8, retaining)
topics: main: author, secondary: report
topics(8, main, id54; secondary, id7; potential, id55)
9. Anaphora Resolution: they resolved as author
state(9, continue)
topics: main: author, secondary: material
topics(9, main, id54; secondary, id61; potential, id62)
10. state(10, continue)
topics: main: author, secondary: site
topics(10, main, id54; secondary, id67; potential, id68)
11. Anaphora Resolution: this resolved as 'discourse bound'; they resolved as author
state(11, retaining)
topics: main: author, secondary: active
topics(11, main, id54; secondary, id71; potential, id72)
12. Anaphora Resolution: they resolved as party
state(12, continue)
topics: main: party, secondary: mistake
topics(12, main, id1; secondary, id78)

4.1 The First Experiment: Anaphora Resolution in Technical Manuals

We downloaded the only freely available corpus annotated with anaphoric relations, i.e. Wolverhampton's Manual Corpus made available by Prof. Ruslan Mitkov on his website. The corpus contains text from Manuals at the following address,

<http://clg.wlv.ac.uk/resources/corpus.html>

Text Type	Referring Exps	Coreferring Exps	Total Words
AIWA	1629	716	6818
ACCESS	1862	513	9381
PANASONIC	1263	537	4829
HINARI	673	292	2878
URBAN	453	81	2222
WINHELP	672	206	2935
CDROM	1944	279	10568
Totals	8496	2624	39631

Table 2. General data of Worlverhampton's coreference annotated corpora

Text Type	Referring Exps % W	Coreferring Exps % RE
AIWA	23.89	43.21
ACCESS	19.84	27.01
PANASONIC	26.15	42.51
HINARI	23.38	29.22
URBAN	20.38	17.88
WINHELP	22.89	27.14
CDROM	18.39	14.24
Means	21.43	30.88

Table 3. Proportion of coreferential expressions to referring expressions

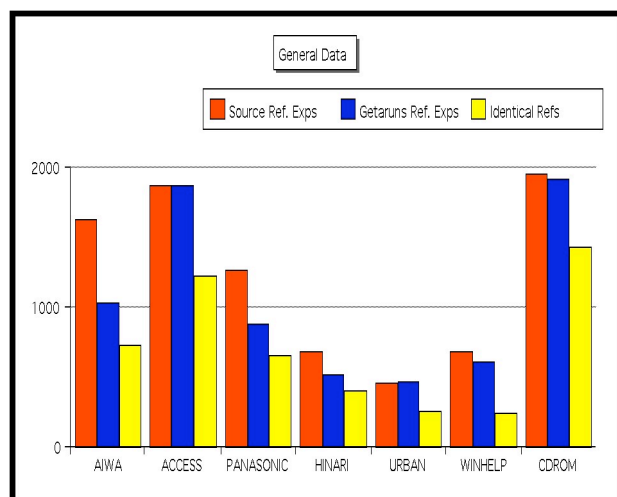


Fig. 3. Comparing GETARUNS output to WMC

We reported in Tab. 2 the general data of the Coreference Corpus. As can be easily noted, there is no direct relationship existing between the number of referring expressions and the number of coreferring expressions. We assume that the higher the number of coreferring expressions in a text the higher is the cohesion achieved. Thus the text identified as CDROM has a very small number of coreferring expressions if compared to the total number of referring expressions. The proportion of referring expressions to words and of coreferring

expressions to referring expressions is reported in percent value in table 3. where the most highly cohesive texts are highlighted in italics; highly non cohesive texts are highlighted in bold:

The final results are reported in the following figure where we plot Precision and Recall for each text and then the comprehensive values.

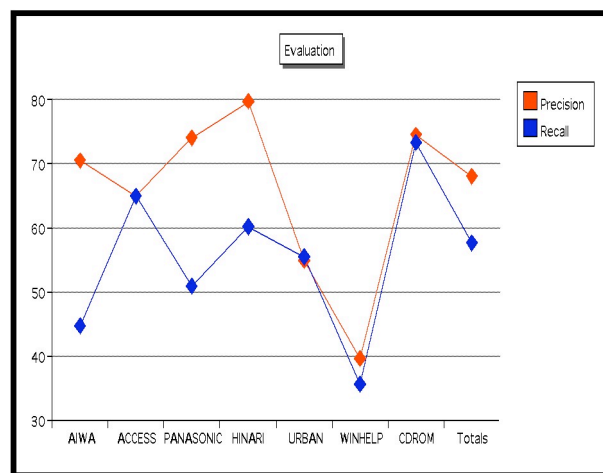


Fig. 4. Precision and Recall for the WMC

4.2 GETARUNS approach to WEB-Q/A

Totally shallow approaches when compared to ours will always be lacking sufficient information for semantic processing at propositional level: in other words, as happens with our "Partial" modality, there will be no possibility of checking for precision in producing predicate-argument structures.

Most systems would use some Word Matching algorithm to count the number of words appearing in both question and the sentence being considered after stripping stopwords: usually two words will match if they share the same morphological root after some stemming has taken place. Most QA systems presented in the literature rely on the classification of words into two classes: function and content words. They don't make use of a Discourse Model where input text has been transformed via a rigorous semantic mapping algorithm: they rather access tagged input text in order to sort best matched words, phrases or sentences according to some scoring function. It is an accepted fact that introducing or increasing the amount of linguistic knowledge over crude IR-based systems will contribute substantial improvements. In particular, systems based on simple Named-Entity identification tasks are too rigid to be able to match phrase relations constraints often involved in a natural language query.

We raise a number of objections to these approaches: first objection is the impossibility to take into account pronominal expressions, their relations and properties as belonging to the antecedent, if no head transformation has taken place during the analysis process.

Another objection comes from the treatment of the Question: it is usually the case that QA systems divide the question to be answered into two parts: the Question

Target represented by the wh- word and the rest of the sentence; otherwise the words making up the yes/no question are taken in their order, and then a match takes place in order to identify most likely answers in relation to the rest/whole of the sentence except for stopwords. However, it is just the semantic relations that need to be captured and not just the words making up the question that matter. Some systems implemented more sophisticated methods (notably [8;9;10]) using syntactic-semantic question analysis. This involves a robust syntactic-semantic parser to analyze the question and candidate answers, and a matcher that combines word- and parse-tree-level information to identify answer passages more precisely.

4.3 A Prototype Q/A system for the web

We experimented our approach over the web using 450 factoid questions from TREC. On a first run the base system only used an off-the-shelf tagger in order to recover main verb from the query. In this way we managed to get 67% correct results, by this meaning that the correct answer was contained in the best five snippets selected by the BOWs system on the output of Google API. However, only 30% of the total correct results had the right snippet ranked in position one.

Then we applied GETARUNS shallow on the best five snippets with the intent of improving the automatic ranking of the system and have the best snippet always position as first possibility. Here below is a figure showing the main components for GETARUNS based analysis.

We will present two examples and discuss them in some detail. The questions are the following ones:

Q: Who was elected president of South Africa in 1994?

A: Nelson Mandela

Q: When was Abraham Lincoln born?

A: Lincoln was born February_12_1809

The answers produced by our system are indicated after each question. Now consider the best five snippets as filtered by the BOWs system:

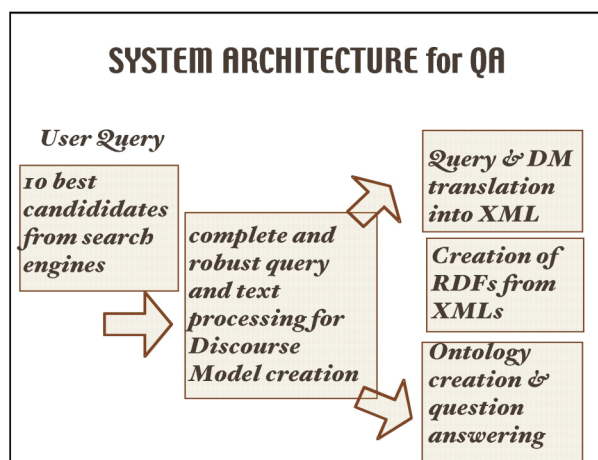


Fig. 5. System Architecture for QA

who/WP was/VBD elected/VBN president/NN of/IN south/JJ africa/NN in/IN 1994/CD

Main keywords: president south africa 1994

Verb roots: elect

Google search: elected president south africa 1994

1.On June 2, 1999, Mbeki, the pragmatic deputy president of South Africa and leader of the African National Congress, was elected president in a landslide, having already assumed many of Mandela's governing responsibilities shortly after Mandela won South Africa's first democratic election in 1994.

2.Washington ? President Bill Clinton announced yesterday a doubling in US assistance South Africa of \$600-million (R2 160-million) over three years, and said his wife Hillary would attend Nelson Mandela's inauguration as the country's first black president.

3.Nelson Mandela, President of the African National Congress (ANC), casting the ballot in his country's first all-race elections, in April 1994 at Ohlange High School near Durban, South Africa.

4.Newly-elected President Nelson Mandela addressing the crowd from a balcony of the Town Hall in Pretoria, South Africa on May 10, 1994.

5.The CDF boycotted talks in King William's Town yesterday called by the South African government and the Transitional Executive Council to smooth the way for the peaceful reincorporation of the homeland into South Africa following the resignation of Oupa Gqozo as president.

Notice snippet n.1 where two presidents are present and two dates are reported for each one: however the relation "president" is only indicated for the wrong one, Mbeki and the system rejects it. The answer is collected from snippet no.4 instead. As a matter of fact, after computing the ADM, the system decides to rerank the snippets and use the contents of snippet 4 for the answer. Now the second question:

when/WRB was/VBD abraham/NN lincoln/NN born/VBN

Main keywords: abraham lincoln

Verb roots: bear

Google search: abraham lincoln born

1. Abraham Lincoln was born in a log cabin in Kentucky to Thomas and Nancy Lincoln.

2. Two months later on February 12, 1809, Abraham Lincoln was born in a one-room log cabin near the Sinking Spring.

3. Abraham Lincoln was born in a log cabin near Hodgenville, Kentucky.

4.Lincoln himself set the date of his birth at feb_12, 1809, though some have attempted to disprove that claim.

5. A. Lincoln (February 12, 1809 April 15, 1865) was the 16/th president of the United States of America.

In this case, snippet n.2 is selected by the system as the one containing the required information to answer the question. In both cases, the answer is built from the ADM, so it is not precisely the case that the snippets are selected for the answer: they are nonetheless reranked to make the answer available.

5. System Evaluation

After running with GETARUNS, the 450 questions recovered the whole of the original correct result 67% from first snippet.

The complete system has been tested with a set of texts derived from newspapers, narrative texts, children stories. The performance is 75% correct. However, updating and tuning of the system is required for each

new text whenever a new semantic relation is introduced by the parser and the semantics does not provide the appropriate mapping. For instance, consider the case of the constituent "holes in the tree", where the syntax produces the appropriate structure but the semantics does not map "holes" as being in a LOCATION semantic relation with "tree". In lack of such a semantic role information a dummy "MODal" will be produced which however will not generate the adequate semantic mapping in the DM and the meaning is lost.

As to the partial system, it has been used for DUC summarization contest, i.e. it has run over approximately 1 million words, including training and test sets, for a number of sentences totalling over 50K. We tested the "Partial" modality with an additional 90,000 words texts taken from the testset made available by DUC 2002 contest. On a preliminary perusal of samples of the results, we calculated 85% Precision on parsing and 70% on semantic mapping. However evaluating full results requires a manually annotated database in which all linguistic properties have been carefully decided by human annotators. In lack of such a database, we are unable to provide precise performance data. The system has also been used for the RTE Challenge and performance was over 60% correct [11].

6. Conclusions

Results reported in the experiment above have been limited to the ability of the system to cope with what has always been regarded as the toughest task for an NLP system to cope with. We have not addressed the problem of question answering for lack of space.

Would it be possible for computers to recognize the layout of a Web page, much in the same manner as a human? Much like the development of the Semantic Web itself, early efforts to integrate natural language technology with the Semantic Web will no doubt be slow and incremental. By weaving natural language into the basic fabric of the Semantic Web, we can begin to create an enormous network of knowledge easily accessible by both machines and humans alike. Furthermore, we believe that natural language querying capabilities will be a key component of any future Semantic Web system. By providing "natural" means for creating and accessing information on the Semantic Web, we can dramatically lower the barrier of entry to the Semantic Web. Natural language support gives users a whole new way of interacting with any information system, and from a knowledge engineering point of view, natural language technology divorces the majority of users from the need to understand formal ontologies. As we have tried to show in the paper, this calls for better NLP tools where a lot of effort has to be put in order to allow for complete and shallow techniques to coalesce smoothly into one single system. GETARUNS represents such a hybrid system and its performance is steadily improving.

In the future we intend to address the problem of using the database of TEs created by our system in answering a more extended set of natural language queries than what has been tried so far.

References

1. Dan Klein and Christopher D. Manning: Accurate Unlexicalized Parsing. ACL, (2003) 423-430
2. D. Lin.: Dependency-based evaluation of MINIPAR. In Proceedings of the Workshop on Evaluation of Parsing Systems at LREC 1998. Granada, Spain, (1998)
3. Sleator, Daniel, and Davy Temperley: "Parsing English with a Link Grammar." Proceedings of IWPT '93, (1993)
4. Delmonte R.: Parsing Preferences and Linguistic Strategies, in *LDV-Forum - Zeitschrift fuer Computerlinguistik und Sprachtechnologie - "Communicating Agents"*, Band 17, 1,2, (2000) 56-73
5. Delmonte R.: Parsing with GETARUN, *Proc.TALN2000, 7^o conference annuel sur le TALN*, Lausanne, (2000) 133-146
6. Delmonte R., D. Bianchi: From Deep to Partial Understanding with GETARUNS, *Proc. ROMAND 2002*, Università Roma2, Roma, (2002) 57-71
7. Boris Katz, Jimmy J. Lin, Sue Felshin: The START Multimedia Information System: Current Technology and Future Directions, In Proceedings of the International Workshop on Multimedia Information Systems (MIS 2002)
8. Hovy, E., U. Hermjakob, & C. Lin.: The Use of External Knowledge in Factoid QA. In E. M. Voorhees & D. K. Harman (eds.), *The Tenth Text Retrieval Conference (TREC 2001)*. (2002) 644-652
9. Litkowski, K. C.: Syntactic Clues and Lexical Resources in Question-Answering. In E. M. Voorhees & D. K. Harman (eds.), *The Ninth Text Retrieval Conference (TREC-9)*. (2001) 157-166
10. Litkowski, K. C.: CL Research Experiments in TREC-10 Question-Answering. In E. M. Voorhees & D. K. Harman (eds.), *The Tenth Text Retrieval Conference (TREC 2001)*. (2002) 122-131
11. Delmonte R., Sara Tonelli, Marco Aldo Piccolino Boniforti, Antonella Bristot, Emanuele Pianta: VENSES – a Linguistically-Based System for Semantic Evaluation, RTE Challenge Workshop, Southampton, PASCAL - European Network of Excellence, (2005) 49-52
12. Delmonte R.: Evaluating GETARUNS Parser with GREVAL Test Suite, Proc. ROMAND - 20th International Conference on Computational Linguistics - COLING, University of Geneva, (2004) 32-41.
13. Bresnan J.(ed.): *The Mental Representation of Grammatical Relations*, MIT Press, Cambridge Mass., 1982)
14. Barwise J., J.M.Gawron, G.Plotkin, S.Tutiya(eds.): *Situation Theory and its Applications*, Vol.2, CSLI Lecture Notes No.26, (1991)