VENSES GetAsk: a System for Hybrid Question Answering And Answer Recovery using Text Entailment

Rodolfo Delmonte

Department of Language Studies and Comparative Cultures
& Department of Computer Science
Ca' Foscari University - Dorsoduro 1075
30123 - VENEZIA (It)
E-mail: delmont@unive.it
Website: project.cgm.unive.it

Abstract. We present a system that couples techniques belonging to Information Extraction and deep linguistic processing for Question Answering. The system presented in the paper has undergone extensive testing and the parser has been trained on available testsuites. The system uses text entailment processing to select best sentences to match with each question. Both sentences and questions need to parsed syntactically and semantically and a logical form has to be produced with predicate argument structures and propositional level analysis. In order to pick the right answer from a set of five, after extracting the best sentence/s from the text, we organized different strategies according to question type and semantic propositional type. The system has access to a wide range of computational lexica, ontologies and datasets to carry out the task: for common sense knowledge we used ConceptNet.

Keywords. syntactic and semantic processing, predicate argument structures,
logical forms, propositional level semantics, semantic disambiguation

1 Introduction

We present a system for question answering that couples statistical processing and deep linguistic analysis, GetAsk, based on GETARUNS, the system for text understanding developed at the Ca' Foscari University of Venice. Like other similar systems (see in particular Bos et al. and Ahn et al.), the architecture of the system is organized as a standard pipeline of interconnected modules: Text or Passage Analysis, Question Analysis, Answer Extraction and Reranking. There is no Document Retrieval phase in our case, since we are supposed to receive texts/passages already selected from a bigger set and supposedly containing the answer to the question.

Both text and question are analysed by our system for text understanding and the output of text analysis is recorded on file in a linear nonrecursive unscoped Logical Form format which is derived directly from the Situation Semantics representation that the system computes. In case of failure of the deep modules, the system still produces a Logical form directly from Dependency structure. Answer extraction and reranking is performed by means of three sequential and incremental filters or sieves:

- at first we use information retrieval methodologies
- the second pass through the text is done by applying semantic similarity measures to the lemmata of sentences selected by the previous filter;
- eventually, we reinforce our previous choices by adding words selected on the basis of Logical Forms as they are computed from question analysis and text analysis.

More details of the whole system in the sections below. We present GETARUNS at first and then the backoff system that runs before and after the deep parser, in order to recover from possible failures; then in section 3 we present the deep method to compute answers directly from the Discourse Model; in section 4 the hybrid version of the system, where we comment in detail on QA4MRE dataset; in section 5 we report some evaluation and we end up with some conclusions.

2 The System VENSES GETARUNS

GETARUNS is organized into three subsystems. Venses is the shallow or partial version of GETARUNS: it is fully bottom-up and is responsible for tagging and chunking and produces a full-fledged syntactic and semantic analysis in case the second system fails. System two is the main deep system: it is organized in two versions. The first version runs fully top-down and the second one on the contrary runs bottom-up. They have access to the same rules which however are taken in a strict top-down order, by the first system; whereas the second system has a bottom-up access to the rules by means of a recursive procedure that is triggered by the current string and information coming from shallow analysis. The output structure is an annotated c-structure which is interpreted by the same Lexical Semantic interpretation module described below. GETARUNS switches to the second bottom-up deep system whenever there is a failure in the top-down parser, or when the sentence to be parsed is longer than 50 tokens. It switches to the "shallow" version in case of failure of the deep system as a whole. Also the "shallow" system produces a semantic representation which is partially coincident with the one produced by the deep system. "Partially" here means that only essential semantic modules are activated: semantic roles assignment; pronominal binding and anaphora resolution; logical form creation. No spatiotemporal resoning is present in the partial system, nor quantifier raising.

The deep system 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 may take place.

The system is based on LFG (Lexical-Functional Grammar) theoretical framework and has a highly interconnected modular structure. The Closed Domain version of the system is a top-down depth-first DCG-based parser written in Prolog Horn Clauses, which uses a strong deterministic policy by means of a lookahead mechanism. A second version of the same set of rules is activated in case of failure, but with a bottomup schema. The output of this second pass is then submitted to the same interpretation module that checks for grammaticality on the basis of lexical subcategorization information made available by any of the currently available computational lexica. Eventually, in case of failure of this second pass, the system derives an interpretation from the "shallow" or partial parse computed as a starting pass, where also tagging takes place, and Head-Dependent structures are built for further use. In fact, the output of this parser is used by the bottom-up deep parser to detect the presence of a verbal constituent while recursively consuming the input string.

The system is divided up into a pipeline of sequential but independent modules which realize the subdivision of a parsing scheme as proposed in LFG theory. We build a c-structure before the f-structure can be projected by unification into a DAG (Direct Acyclic Graph) – however we map c-structures to DAG using Prolog unification. In this sense we try to apply in a given sequence phrase-structure rules as they are ordered in the grammar: whenever a syntactic constituent is successfully built, it is checked for semantic consistency. In case the governing predicate expects obligatory arguments to be lexically realized they will be searched and checked for uniqueness and coherence as LFG grammaticality principles require.

Syntactic and semantic information is accessed and used as soon as possible: in particular, both categorial and subcategorization information attached to predicates in the lexicon is extracted as soon as the main predicate is processed, be it adjective, noun or verb, and is used to subsequently restrict the number of possible structures to be built. Adjuncts are computed by semantic compatibility tests on the basis of selectional restrictions of main predicates and adjuncts heads. The subdivision of arguments and adjuncts is guided by available lexica, and ambiguity is solved by frequency counts associated to Verb or Noun argument/adjunct taken from Penn Treebank.

The grammar is equipped with a core lexicon containing most frequent 5000 fully specified inflected word forms where each entry is followed by its lemma and a list of morphological features, organised in the form of attribute-value pairs. However, morphological analysers for English are also available with big root dictionaries (25,000 for English) which only provide for syntactic subcategorization, though. In addition to that there are all lexical form provided by a fully revised version of COMLEX, and 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 are used to generate a subcategorization schemes with an aspectual and semantic class associated to it – however no selctional restrictions can reasonably be formulated on arguments of predicates. Semantic inherent features for Out of Vocabulary Words, be they nouns, verbs, adjectives or adverbs, are provided by a fully revised version of WordNet - plus EuroWordnet, with a number of additions coming from additional specialized semantic fields like computer, economics, and advertising - in which we used 75 semantic classes similar to those provided by CoreLex.

When each sentence is parsed, tense aspect and temporal adjuncts are accessed to build the basic temporal interpretation to be used by the temporal reasoner. Eventually two important modules are fired: Quantifier Raising and Pronominal Binding. QR is computed on f-structure which is represented internally as a DAG. It may introduce a pair of functional components: an operator where the quantifier can be raised, and a pool containing the associated variable where the quantifier is actually placed in the f-structure representation. This information may then be used by the following higher system to inspect quantifier scope. Pronominal binding is carried out at first at sentence internal level. DAGs will be searched for binding domains and antecedents matched to the pronouns if any to produce a list of possible bindings. Best candidates will then be chosen. After these modules have been successfully fired, the f-structure is completed and cannot undergo further changes.

2.1 The Upper Module

GETARUNS, has a common (for both versions of the deep system) 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 sub-modules which take care of Spatio-Temporal Reasoning, Discourse Level Anaphora Resolution, and other subsidiary processes like Topic Hierarchy which cooperate to find the most probable antecedent of coreferring and cospecifying referential expressions when creating semantic individuals. These are then asserted in the Discourse Model (hence the DM), which is then the sole knowledge representation used to solve nominal coreference. The system uses two resolution submodules which work in sequence: they constitute independent modules and allow no backtracking. The first one is fired whenever a free sentence external pronoun is spotted; the second one takes the results of the first sub-module and checks for nominal anaphora. They have access to all data structures contemporarily and pass the resolved pair, anaphor-antecedent to the following modules. Semantic Mapping is performed in two steps: at first a Logical Form is produced which is a structural mapping from DAGs onto unscoped well-formed formulas. These are then turned into situational semantics informational units, infons which may become facts or sits. Each unit has a relation, a list of arguments which in our case receive their semantic roles from lower processing – a polarity, a temporal and a spatial location index.

2.2 Incremental Shallow-to-Deep Parsing

The so-called shallow or partial module, is rather generic. As in most shallow parsers, we use a sequence or cascade of transducers: however, in our approach, since we intend to recover sentence level structure, the process goes from partial parses to full parses. Sentence and then clause level is crucially responsible for the right assignment of arguments and adjuncts to a governing predicate head. This is clearly paramount in our scheme which aims at recovering predicate-argument structures, besides performing a compositional semantic translation of each semantically headed constituent.

3 Hybrid Question-Answering

3.1 State of the art and our approach

When compared to our approach, totally shallow IR/IE approaches 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 that counts the number of words that appear in both the 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 match words, phrases or sentences according to some matching scoring function (see the TREC QA series on NIST website).

It is also common knowledge the fact that only by 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.

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.

Second objection is the use of grammatical function labels, like SUBJ/OBJects without an evaluation of their relevance in the utterance structure: higher level or main clause SUBJ/OBJects are more important than other SUBJects. In addition, there is no attempt at semantic role assignment which would come from a basic syntactic/semantic tagging of governing verbs: a distinction into movement verbs, communication verbs, copulative verbs, psychic verbs etc. would suffice to assign semantic roles to main arguments if present.

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 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 only the words making up the question that matter. Some system implemented more sophisticated methods (notably Hovy et al.; Litkowski; Bos et al.): syntactic-semantic question analysis. This involves a robust syntactic-semantic parser to analyse the question and candidate answers, and a matcher that combines word- and parse-tree-level information to identify answer passages more precisely.

More closely related to our approach are two systems that we shall comment here below. The first one is presented in Dan Moldovan et al. and is the LCC system called PowerAnswer. As the authors comment, it obtained a confidence weighted score of 0.85% on a dataset of 500 questions at TREC QA 2002. In their introduction the authors present the component of their system combines syntactic, semantic, lexical and world knowledge information sources (Moldovan et al.). Questions and relevant document paragraphs are transformed into logical forms that together with world knowledge axioms extracted from WordNet glosses are fed to a logic prover (Moldovan et al.). In order to allow the syntactic parser to work in a reasonable time they feed it with only relevant text excerpts that have been previously extracted by a summarization system. They also do coreference resolution by equating definite expressions with their antecedent in case it is a personal proper name; but also other more complex forms of coreference involving indefinite and definite noun phrase and pronoun coreference have been implemented (Moldovan et al.). The output LF is then passed to a theorem logic prover that checks the result.

The main difference with our approach lies in the fact that they produce a shallow syntactic analysis and only after that they start introducing logic constraints. On the contrary, we use all possible constraints at the moment of semantic mapping from syntactic structure which in our case is never shallow – not just considering surface structure but introducing all relevant missing and implicit arguments.

If we look at the other approach presented by Barker et al. 2007, we see that the same surface level syntactic – dependency-based – analysis is produced before mapping into logical forms. Their system introduces special axioms to take care of domain world knowledge, and some general semantic definition, as for instance, translating plural noun phrases into sets. The output LF is then passed on to a reasoner that checks the result.

3.2 Our approach

We have a passage ranking component that takes a query and a set of documents, it extracts sentences, and assigns a score to them. This is done by two passages over each

text, where on a first passage, after lowcasing and lemmatizing all words in text and query we retain information related to sentences where:
- we count the number of non-stopword query word tokens (as opposed to types) present in the sentence that are positive to an identity match, and the result is not an empty set;

On a second pass, on the contrary, we keep the original orthography and take care of words beginning with uppercase letters, and we count:
- all words that match semantically, by accessing WordNet and other computational lexica – we use Sumo-Milo and FrameNet.

The non empty matching results are then passed to another important filter that takes Logical Form of the query and looks for heads and predicates of predicate-argument structures contained there. The final score obtained is the sum of the previous computation and the last one, where we impose the presence of the most relevant lemmas in the choice of the best candidate sentence.

Logical Forms are derived from DAGs of f-structure sentence level representation and are simplified in order to be useful for the question answering task. In particular, we come up with a non-recursive linear representation at propositional level where we introduce prefixes for each semantic head which are very close to DRS-conditions:
- PRED, QUANT, CARD, ARG, MOD, ADJ, FOC

where Foc contains the question type derived from a mapping of each wh- word, together with its possible nominal or adjectival head and a restricted set of semantic general classes, like MEASURE, MANNER, QUANTITY, REASON etc.

The text representation is made in the form of Discourse Model which is simplified before matching takes place. In particular, we compose two types of semantic structures from the list of facts:
- an event structure for each governing predicate which includes the arguments in their literal form and their semantic indices, together with the polarity and the two spatiotemporal indices;
- an enriched version of the fact associated to each entity in the DM, which includes knowledge of the world (synset and definition) retrieved in one of the ontologies and computational lexica available;
- a relational representation for each relation present in the DM that associates properties to entities and relations, including discourse markers at propositional level, attributes, modifiers, partonimy, generic unmarked relations (OF relation) etc.

This level of representation is used to match possible answers with the chosen sentence, thus trying to select the most appropriate answer cadidates.

4 The *QA4MRE Main Task* dataset

In the *QA4MRE* dataset for English, we go from simple factoid questions to highly complex and sometimes hardly understandable questions. In between, in some cases, the correct answer made available is not a direct answer but requires some reasoning to be in place in order for the system to select it. In some cases there are more answer right, while in other cases none of the answers is correct.

As to resources used to answer questions, we found it very important to access the commonsense reasoning repositoire called CONCEPTNET, as made available by MIT AI laboratory. This is done whenever the similarity algorithm has attemped all possible semantic inferencing steps and has reached a failure. Eventually, access to commonsense reasoning is produced in order to fill in the gap of some intermediate reasoning link or step which needs to be restored in order for the appropriate answer to be selected. We will comment specific cases in the sections below.

What the system does is to use Logical Forms in order to produce matches between Question appropriately turned into the corresponding prospective Answer, and sentences contained in the text. Whenever matches are found a score is generated which allows the system to grade best sentence candidates to be considered in the second part of the analysis, when the best answer is to be chosen from the set of answers made available in the dataset.

At first we produce a surface level identity match of the actual words contained in question and candidate sentence using the typical Information Retrieval approach: we go through each word and skip stop words. If we don't find a match, we try with lemmaized version of question and text sentences. This first pass through the text produces a score which is then passed to the second level matching mechanism that relies on Semantics. It is worth noting, that in this second level, all unexpressed linguistic elements are placed in their required position by Logical Form constraints that need, for instance, SUBJects to be in place before a complete Predicate-Argument structure is built. We also recover antecedents of pronominal expressions as they have been computed by the Anaphora Resolution algorithm included in our system.

Matches are produced by doing Identity match at first, between Heads that constitute the Predicate-Argument structure contained in the LF of the Question and the candidate sentence. We are using a mechanism which is derived directly from our previous work on RTE, which not only allows us to detect mismatches but also contradictions thus rejecting the candidate with a low score.

All similarity matches are produced by inferencing with WordNet and other similar resources, also on the basis of semantic general tags, like the ones introduced by SUMO-MILO.

We will only comment on Text 13 in details: this text is one of the most difficult to answer – if not the most difficult. Difficulties arise basically due to the need to produce both anaphora and coreference resultion links between entities and events mentioned in succession. Questions "easy" to answer are those that "literally" coincide with the semantic contents of one sentence in the text: that is, the predicate-argument structure coincides with the one of the question, and the entities mentioned are semantically identical or very similar to the one contained in the question. As will be clear from the comments below, there are only three questions over 18 which can be regarded “easy” to answer. This is also a text that contains 3 “why” questions and one “How many” question: these are usually regarded most difficult questions to answer.

We report here below a long excerpt from the first part of the text, and then make short references to the remaining part. For each question of the 18 proposed, we list the answers and then make comments on the right choice and the difficulties inherent in finding it. Here is the excerpt:

> The appointment of a former top executive of a major U.S. pharmaceutical company and major Republican contributor as President George W. Bush's global AIDS co-ordinator has stunned and outraged AIDS experts and activists. Bush's choice of former Eli Lilly & Co. boss Randall Tobias was announced at the White House on July 1, just a few days before Bush's first trip as president to Africa. The U.S. Senate must confirm the nomination. Tobias, who retired from Lilly in 1998 and more recently has served as vice chairman of AT&T, where he also worked before going to Lilly in the early 1990s, is supposed to receive the rank of ambassador and report to Secretary of State Colin Powell, a major force behind a five-year, 15-billion-dollar anti-AIDS initiative - called the "Emergency Program" - first proposed by Bush last January and approved by Congress in a somewhat amended form in May. Implementation of that initiative, which is targeted at 12 sub-Saharan African and two Caribbean countries, will be Tobias' first responsibility, according to Bush. "Randy Tobias has a mandate directly from me to get our AIDS initiative up and running as soon as possible," he said. Surreal Appointment Prof. Jeffrey Sachs, head of Columbia University's Earth Institute and a special adviser to UN Secretary General Kofi Annan on the AIDS crisis, called the appointment "surreal" and continued that "This is an emergency that requires someone who's worked in the field and knows it thoroughly. We don't need someone who raises all sorts of questions about commitment and agenda." Advocacy groups called for senators to closely scrutinize Tobias' credentials and philosophy and determine whether, given his past ties to the industry, he will be able to fight on behalf of the millions of poor HIV/AIDS victims in desperate need of cheap anti-retroviral drugs in the face of opposition from the major western pharmaceutical companies, often referred to as Big Pharma. "This decision is another deeply disturbing sign that the President may not be prepared to fulfill his pledge to take emergency action on AIDS," noted Paul Zeitz, executive director of the Global AIDS Alliance. "It raises serious questions of conflict of interest and the priorities of the White House." "Both the people of Africa and the people of the United States will lose if the president's AIDS initiative fails to use the lowest-cost, generic medications," Zeitz said, noting that the pharmaceutical companies have successfully pressed the Bush administration to go back on an earlier pledge to carve out an exception in international patent laws that would enable needy countries to import generic anti-AIDS drugs.

Quest.: 1, What is the main objective of the Emergency Program ?

ans(1, to make anti-retroviral drugs available to the poor), ans(2, to use the lowest-cost generic medications), ans(3, to change the international patent laws), ans(4, to import life-saving drugs), ans(5, none of the above)

The best right answer is answer 1 and can be found in the text reported above, further down, four sentences below after the reference to Tobias. Also answer 2 is correct and can be found in a comment at the end of the excerpt. The problem is that this can only happen in case all anaphora and coreference resolution steps have been correctly performed. At the beginning we are told that Tobias is responsible for the implementation of the "Emergency Program" which is then mentioned as "that initiative". The same program is coreferred to by Annan as "this emergency". Eventually, the goals of the initiative are introduced in a following sentence, where "Tobias' credentials" will be scrutinized to determine whether "he will be able to fight on behalf of the millions of poor HIV/AIDS victims in desperate need of cheap anti-retroviral drugs".

Quest.: 2, Why were AIDS activists not happy with Randall Tobias being appointed as global AIDS co-ordinator ?

ans(1, because he was the head of Columbia University), ans(2, because he was supposed to favour the pharmaceutical industries), ans(3, because he lived in Caribbean countries), ans(4, because he was a person with great acumen), ans(5, none of the above)

Question two is best answered by answer 2. and is found in the same piece of text reported above. Here again we may note that the answer uses a different wording from what can be found in the text with the same meaning: "pharmaceutical industries" rather than "pharmaceutical companies". However understanding that the portion of selected text is actually talking about AIDS activists unhappy with Randall Tobias appointed as global AIDS coordinator is not an easy task.

Quest.: 3, Why is Randall Tobias supposed to receive the rank of ambassador ?

ans(1, because he was a major Republican contributor), ans(2, because he was a former top executive of a major U.S. pharmaceutical company), ans(3, because he retired from Lilly), ans(4, because he was vice chairman of A&T), ans(5, none of the above)

Question 3 doesn't have an answer, so answer 5 would be the best choice.

Quest.: 4, Has Randall Tobias been confirmed as President George W. Bush's global AIDS co-ordinator ?

ans(1, Yes, a few days before Bush's first trip as president to Africa), ans(2, Not yet), ans(3, Yes, on July 1), ans(4, Yes, last January), ans(5, none of the above)

Here there is only one possible answer, and it is answer 2. This is derivable from this excerpt, where we see that there has been an "announcement" of nomination but it hasn't been confirmed yet:

> The appointment of a former top executive of a major U.S. pharmaceutical company and major Republican contributor as President George W. Bush's global AIDS co-ordinator … Bush's choice of former Eli Lilly & Co. boss Randall Tobias was announced at the White House on July 1, just a few days before Bush's first trip as president to Africa. The U.S. Senate must confirm the nomination.

In order to be able to associate "Not yet" to the second sentence, the system needs to corefer "Nomination" to "Bush's best choice", and link the latter to "Appointment" in the previous sentence. In other words, the text reports an "appointment" then a "choice" and eventually a "nomination". If appointment and nomination are perfect synonyms, "choice" isn't included in any synset related to them. The link between choice and nomination is then missing.

Quest.: 5, What does the author of the book "The End of Poverty" think about the appointment of Randall Tobias ?

ans(1, he defines it as important), ans(2, he defines it as successful), ans(3, he defines it as serious), ans(4, he defines it as surreal), ans(5, none of the above)

No author of a book is mentioned in the text so the answer has to be answer 5.

Quest.: 6, Who will be in charge of carrying out effectively the "Emergency Plan" ?

ans(1, George W. Bush), ans(2, the former chief executive officer of Eli Lilly & Co), ans(3, Secretary of State Colin Powell), ans(4, the head of Columbia University), ans(5, none of the above)

The right answer is answer 2, with a long description of properties which are again referring to Tobias. However "carrying out effectively" is to be understood as a paraphrase of "implementation", which is what we find in the text. The synonym link appears in WordNet, but coreference between a noun "implementation" and the verb "carry out" is not easy to perform.

Quest.: 7, What does Jeffrey Sachs think about the appointment of Randall Tobias ?

ans(1, he defines it as important), ans(2, he defines it as successful), ans(3, he defines it as serious), ans(4, he defines it as surreal), ans(5, none of the above)

Right answer is answer no. 4, where the appointment is defined as "surreal". This is the only easy question to answer. The problem in this case is constituted by the need to use a coreferring singular definite nominal "appointment" that needs to be linked to the previous mention, beginning of the text, where however its subject is only indirectly referred to Tobias:

> Prof. Jeffrey Sachs, head of Columbia University's Earth Institute and a special adviser to UN Secretary General Kofi Annan on the AIDS crisis, called the appointment "surreal"…

Quest.: 8, What types of drug are used by the U.S. Administration for the Emergency Program ?

ans(1, generics), ans(2, it is to be decided), ans(3, brand-name anti-viral medicines), ans(4, triple combinations of anti-retroviral drugs), ans(5, none of the above)

Here the right answer is answer 2. Again the answer is not directly available and needs some inference to be fired from the following excerpt:

> While the administration has suggested it will use generics in the Emergency Program, it has not been made a formal decision.

Quest.: 9, How many countries are included in the Emergency Program ?

ans(1, 12), ans(2, 2), ans(3, 18), ans(4, 10), ans(5, none of the above)

None of the above is the right answer, as can be gathered from the first excerpt reported above. Of course in order to properly understand the content of the question and pair it with the right piece of text, some inference is needed. The question says "included in the Emergency Program", and the text says "Implementation of that initiative, which is targeted to..." where “targeted to” is followed by the countries.

Quest.: 10, What is a strong characteristic of Randall Tobias ?

ans(1, his experience with AIDS), ans(2, his background in public health), ans(3, his experience with working in poor countries), ans(4, his contacts with the World Trade Organization (WTO)), ans(5, none of the above)

Here the right answer is answer 5, "none of the above". This is again difficult to get.

Quest.: 11, Who was the adviser of the Ghanaian diplomat ?

ans(1, Randall Tobias), ans(2, Jeffrey Sachs), ans(3, Colin Powell), ans(4, George W. Bush), ans(5, none of the above)

As before, the right answer is no. 5, "none of the above". In the text there is no reference to Ghanian diplomats.

Quest.: 12, Who was the adviser of Kofi Annan ?

ans(1, Randall Tobias), ans(2, Jeffrey Sachs), ans(3, Colin Powell), ans(4, George W. Bush), ans(5, None of the above)

Here the right answer is no. 2, Jeffrey Sachs. This is the second easy question to answer.

Quest.: 13, What is Randall Tobias reputation ?

ans(1, he is a down-to-earth business person), ans(2, he is incomprehensible), ans(3, he is an impoverished man), ans(4, he is a man of philosophy), ans(5, None of the above)

Right answer is no. 1. The adjective qualifying the property of being a "business person", is however different in the question, from what is found in the text. In the question we have "down-to-earth" and in the text we have "a no-nonsense": no synonyms are available.

Quest.: 14, What is Randall Tobias' reputation ?

ans(1, he is a no-nonsense businessman), ans(2, he is incomprehensible), ans(3, he is an impoverished man), ans(4, he is a man of philosophy), ans(5, none of the above)

Here on the contrary, the adjective used in the aswers is the same that appears in the text, and is contained in answer no. 1. So this is the third easy answer to get.

Quest.: 15, Where were the agreements on international patent law signed ?

ans(1, at the World Trade Organization meeting in Doha), ans(2, at Big Pharma), ans(3, at the office of Management and Budget), ans(4, at the Health Global Access project meeting), ans(5, none of the above)

The right answer is no. 5, "none of the above". In the text there is no spatial location associated to the event of “signing of agreements”.

Quest.: 16, Why is Big Pharma considered the major organization responsible for contributing to the Global Fund ?

ans(1, because Big Pharma will provide $200 million), ans(2, because Big Pharma is against the Emergency program), ans(3, because Big Pharma produces drugs in India , Thailand and Brazil), ans(4, because Big Pharma wants to import generic anti-AIDS and other life-saving drugs), ans(5, none of the above)

The right answer is no. 5, because Big Pharma is not contributing to the Global Fund. On the contrary, we know from text that it is the "major culprit behind the administration's niggardliness towards the Fund". But obviously, making negative decisions, or finding the contrary of what is being asserted is very difficult.

Quest.: 17, What is the annual US contribution to the Global Fund to fight AIDS ?

ans(1, $200 million), ans(2, $1 billion), ans(3, $20 million), ans(4, $2 billion), ans(5, None of the above)

Here the information needs to be badly filtered and inferences fired. The sentence containing the answer is the following one:

> Although Congress has authorized an annual contribution of up to $1 billion for the Fund - which is already fast running out of money - the administration has said it intends to provide only $200 million a year.

The answer in this case is again "none of the above" and it is hard to compute from the text. In the extracted sentence, we can see that neither the concessive headed by "although", nor the main clause constitute a factual assertion. Since that is what is required by the question, the answer is left unsatisfied and unanswered.

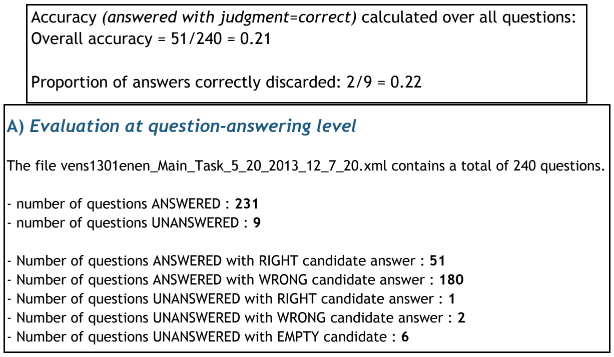
Quest.: 18, What are activists most concerned about ?

ans(1, about statistics of the AIDS toll in Africa), ans(2, about importing generic anti-AIDS drugs), ans(3, about the International AIDS Trust), ans(4, about the World Trade Organization), ans(5, None of the above)

The question uses a superlative "most concerned" which only pairs with the last of three questions posed by activists on Tobias nomination. The text contains the expression "particularly worried" which should be understood as synonymous to the previous adjectives. But then the object does not match any of the possible answer, and so again the right answer is no. 5.

5 Evaluation

As said above, out system doesn't have to go through a training phase. This is positive on the one side but it could become negative in case some unforseen and unpredictable linguistic problem arises in the analysis of either the text or the questions. Results obtained in the final run are not very satisfactory. This is due to difficulties in analyzing some of the texts; but also in some cases to types of questions which were not understood by the system. For this reason we decided to produce a "Late Run", one week after.
At first we report here below results of the analysis of the regular first run.



Accuracy *(answered with judgment=correct)* calculated over all questions:
Overall accuracy = 51/240 = 0.21

Proportion of answers correctly discarded: 2/9 = 0.22

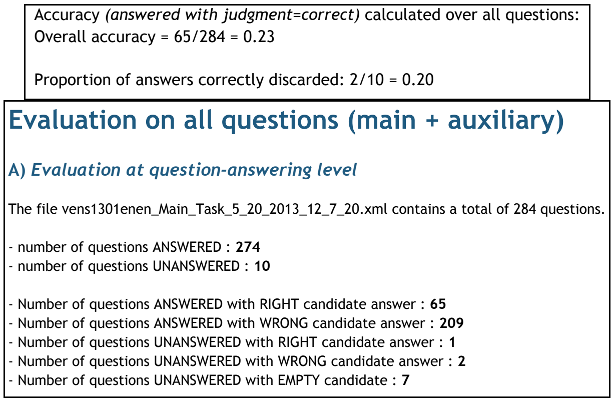
A) ***Evaluation at question-answering level***

The file vens1301enen_Main_Task_5_20_2013_12_7_20.xml contains a total of 240 questions.

- number of questions ANSWERED : **231**
- number of questions UNANSWERED : **9**

- Number of questions ANSWERED with RIGHT candidate answer : **51**
- Number of questions ANSWERED with WRONG candidate answer : **180**
- Number of questions UNANSWERED with RIGHT candidate answer : **1**
- Number of questions UNANSWERED with WRONG candidate answer : **2**
- Number of questions UNANSWERED with EMPTY candidate : **6**

Table 1. Evaluation on Main Questions



Accuracy *(answered with judgment=correct)* calculated over all questions:
Overall accuracy = 65/284 = 0.23

Proportion of answers correctly discarded: 2/10 = 0.20

Evaluation on all questions (main + auxiliary)

A) *Evaluation at question-answering level*

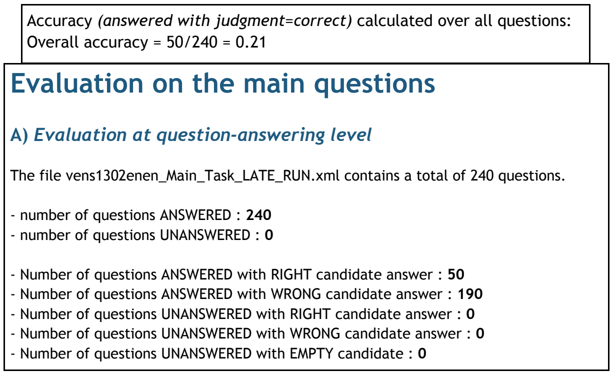
The file vens1301enen_Main_Task_5_20_2013_12_7_20.xml contains a total of 284 questions.

- number of questions ANSWERED : **274**
- number of questions UNANSWERED : **10**

- Number of questions ANSWERED with RIGHT candidate answer : **65**
- Number of questions ANSWERED with WRONG candidate answer : **209**
- Number of questions UNANSWERED with RIGHT candidate answer : **1**
- Number of questions UNANSWERED with WRONG candidate answer : **2**
- Number of questions UNANSWERED with EMPTY candidate : **7**

Table 2. Evaluation on All Questions

And here below we report results of the “LATE runs” which after a first one not remarkably better, except for the fact that the system managed to answer all questions, in the second Late Run reported in Table 5 we see a noticeable recovery.



Accuracy *(answered with judgment=correct)* calculated over all questions:
Overall accuracy = 50/240 = 0.21

Evaluation on the main questions

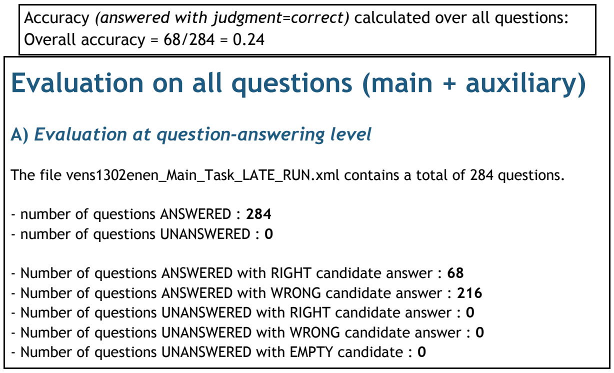
A) *Evaluation at question-answering level*

The file vens1302enen_Main_Task_LATE_RUN.xml contains a total of 240 questions.

- number of questions ANSWERED : **240**
- number of questions UNANSWERED : **0**

- Number of questions ANSWERED with RIGHT candidate answer : **50**
- Number of questions ANSWERED with WRONG candidate answer : **190**
- Number of questions UNANSWERED with RIGHT candidate answer : **0**
- Number of questions UNANSWERED with WRONG candidate answer : **0**
- Number of questions UNANSWERED with EMPTY candidate : **0**

Table 3. Evaluation on Main Questions for Late Run



Accuracy *(answered with judgment=correct)* calculated over all questions:
Overall accuracy = 68/284 = 0.24

Evaluation on all questions (main + auxiliary)

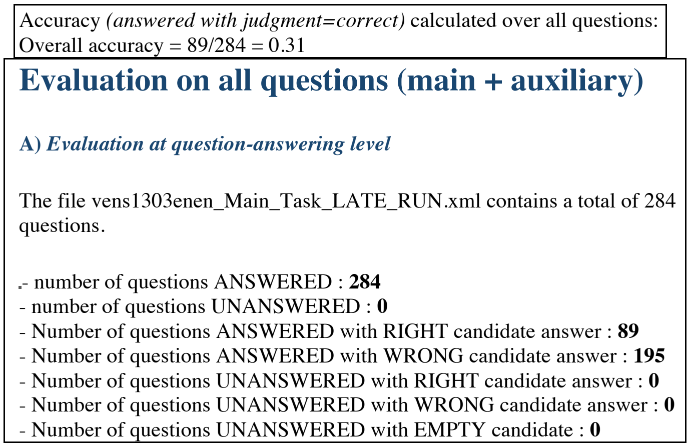
A) ***Evaluation at question-answering level***

The file vens1302enen_Main_Task_LATE_RUN.xml contains a total of 284 questions.

- number of questions ANSWERED : **284**
- number of questions UNANSWERED : **0**

- Number of questions ANSWERED with RIGHT candidate answer : **68**
- Number of questions ANSWERED with WRONG candidate answer : **216**
- Number of questions UNANSWERED with RIGHT candidate answer : **0**
- Number of questions UNANSWERED with WRONG candidate answer : **0**
- Number of questions UNANSWERED with EMPTY candidate : **0**

Table 4. Evaluation on All Questions for Late Run



Accuracy *(answered with judgment=correct)* calculated over all questions:
Overall accuracy = 89/284 = 0.31

Evaluation on all questions (main + auxiliary)

A) ***Evaluation at question-answering level***

The file vens1303enen_Main_Task_LATE_RUN.xml contains a total of 284 questions.

.- number of questions ANSWERED : **284**
- number of questions UNANSWERED : **0**
- Number of questions ANSWERED with RIGHT candidate answer : **89**
- Number of questions ANSWERED with WRONG candidate answer : **195**
- Number of questions UNANSWERED with RIGHT candidate answer : **0**
- Number of questions UNANSWERED with WRONG candidate answer : **0**
- Number of questions UNANSWERED with EMPTY candidate : **0**

Table 5. Resubmission of Evaluation on All Questions for Late Run

Here below we show full statistics of the evaluation:

Overall c@1 measure = (89+0(89/284))/284 = 0.31

Overall c@1 per topic:
c@1 topic t_id '1' = (19+0(19/60))/60 = 0.32
c@1 topic t_id '2' = (32+0(32/78))/78 = 0.41
c@1 topic t_id '3' = (12+0(12/74))/74 = 0.16
c@1 topic t_id '4' = (26+0(26/72))/72 = 0.36

Median: **0.29 -** ***Average:*** **0.31 -** ***Standard Deviation:*** **0.13 -calculated** ***over c@1 of all 16 reading tests***

Topic t_id = '1' - Alzheimer
Median: 0.27 - *Average:* 0.32 - *Standard Deviation:* 0.10 -calculated *over the c@1 of the four reading tests*
- c@1 measure for reading-test r_id '1' = (4+0(4/15))/15 = 0.27
- c@1 measure for reading-test r_id '2' = (7+0(7/15))/15 = 0.47
- c@1 measure for reading-test r_id '3' = (4+0(4/15))/15 = 0.27
- c@1 measure for reading-test r_id '4' = (4+0(4/15))/15 = 0.27

Topic t_id = '2' - Music and society
Median: 0.41 - *Average:* 0.41 - *Standard Deviation:* 0.11 -calculated *over the c@1 of the four reading tests*
- c@1 measure for reading-test r_id '5' = (7+0(7/20))/20 = 0.35
- c@1 measure for reading-test r_id '6' = (10+0(10/19))/19 = 0.53
- c@1 measure for reading-test r_id '7' = (6+0(6/20))/20 = 0.30
- c@1 measure for reading-test r_id '8' = (9+0(9/19))/19 = 0.47

Topic t_id = '3' - Climate Change
Median: 0.14 - *Average:* 0.16 - *Standard Deviation:* 0.07 -calculated *over the c@1 of the four reading tests*
- c@1 measure for reading-test r_id '9' = (2+0(2/18))/18 = 0.11
- c@1 measure for reading-test r_id '10' = (2+0(2/18))/18 = 0.11
- c@1 measure for reading-test r_id '11' = (3+0(3/18))/18 = 0.17
- c@1 measure for reading-test r_id '12' = (5+0(5/20))/20 = 0.25

Topic t_id = '4' - AIDS
Median: 0.36 - *Average:* 0.36 - *Standard Deviation:* 0.07 -calculated *over the c@1 of the four reading tests*
- c@1 measure for reading-test r_id '13' = (6+0(6/18))/18 = 0.33
- c@1 measure for reading-test r_id '14' = (5+0(5/18))/18 = 0.28
- c@1 measure for reading-test r_id '15' = (8+0(8/18))/18 = 0.44
- c@1 measure for reading-test r_id '16' = (7+0(7/18))/18 = 0.39

6 Conclusions

Eventually, the evaluation of the system on the test set is not satisfactory. But this is certainly due to the intrinsic difficulty of the dataset and the way in which questions have been formulated. Our system does both anaphora and coreference resolution and subsequently should be able to allow for long distance coreference. But clearly, the level of accuracy of these two processes is fairly low – differently from pronominal binding which averages 75% accuracy. What we actually must admit is that in order to find the correct answer, the contribution of the Logical Form and semantic Discourse Model is limited to a 30% improvement over an approach in which structural information plays no role whatsoever. In general, BOWs approach is totally inefficient and produces confusing results when the selection of the right answer has to be performed solely on the basis of content word identity match. When lemmatization is added there are improvements but they are not very significant, and this is due to the fact that scoring the best candidate on the basis of word identity match is not enough to distinguish relevant from irrelevant linguistic material. This happens even when we compute on deep rather than surface level analysis. Semantic similarity matches are worked out on the basis of available resources, which however in many case is not sufficient.

We also considered very important the need to distinguish different types of questions, not only on the basis of the question word or question NP, but also and foremost in case the overall question structure requires specific semantic processing to be in place. But the task has been made much harder by the presence of null or negative answers: they are represented by the option no.5 “none of the above”. In order for the system to choose this option, quantitative evaluations should be available that would allow to use graded scales or thresholds to prevent it from accepting approximate solutions. This is not always feasible and our system has not been tuned yet to check for a fine-grained level of semantic consistency. This is going to be our improvements for the future.

7 References

[Bos et al., 2007] Johan Bos, J. R. Curran, E. Guzzetti, The Pronto QA system at TREC-2007: harvesting hyponyms, using nominalisation patterns, and computing answer cardinality. In: E. M. Voorhees, L. P. Buckland (eds): The Sixteenth Text RETrieval Conference, TREC 2007, pp 726–732, Gaitersburg, MD, 2007.

[Ahn et al, 2005], K. Ahn, J. Bos, J. R. Curran, D. Kor, M. Nissim, B. Webber, Question Answering with QED at TREC-2005. In: E. M. Voorhees, L. P. Buckland (eds): The Fourteenth Text REtrieval Conference, TREC 2005, Gaitersburg, MD, 2005.

[Bresnan, 2000] Joan Bresnan, Lexical-Functional Syntax, Blackwell, 2000.

[Delmonte R., 2007] Rodolfo Delmonte, *Computational Linguistic Text Processing - Logical Form, Semantic Interpretation, Discourse Relations and Question Answering*, Nova Science Publishers, New York, 2007.

[Delmonte R., 2009] Rodolfo Delmonte, *Computational Linguistic Text Processing - Lexicon, Grammar, Parsing and Anaphora Resolution*, Nova Science Publishers, New York, 2009.

[Bos & Delmonte, 2008] Johan Bos & Rodolfo Delmonte (eds.), Semantics in Text Processing (STEP), Research in Computational Semantics, Vol.1, College Publications, London, 2008.

[Fellbaum, 1998] Christiane Fellbaum, (ed.) WordNet: An Electronic Lexical Database. MIT Press, Cambridge MA, 1998.

ComLex:- http://nlp.cs.nyu.edu/comlex.

CoreLex:-http://www.cs.brandeis.edu/~paulb/ CoreLex/ corelex.html

EuroWordNet:- http://www.illc.uva.nl/EuroWordNet/

WordNet-Affect:- http://wndomains.fbk.eu/wnaffect.html

[Schwitter et al. 2000] Schwitter R., D. Mollà, R. Fournier & M. Hess, 2000. Answer Extraction: Towards better Evaluations of NLP Systems. *In Proc. Works. Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, Seattle, 20-27.

[Hirschman et al. 1999] Hirschman, L. Marc Light, Eric Breck, & J. D. Buger. Deep

Read: A reading comprehension system. In *Proc. A CL '99*.University of Maryland.

[Hovy et al. 2002] Hovy, E., U. Hermjakob, & C. Lin. (2002a). The Use of External Knowledge in Factoid QA. In E. M. Voorhees & D. K. Harman (eds.), *The Tenth Text Retrieval Conference (TREC 2001)*. 644-652.

[Litkowski, 2001] Litkowski, K. C. (2001). Syntactic Clues and Lexical Resources in Question-Answering. In E. M. Voorhees & D. K. Harman (eds.), *The Ninth Text Retrieval Conference (TREC-9)*. 157-166.