Can We Derive General World Knowledge from Texts?*

Lenhart Schubert
University of Rochester
Rochester, NY 14627-0226
schubert@cs.rochester.edu

ABSTRACT

As one attack on the "knowledge acquisition bottleneck", we are attempting to exploit a largely untapped source of general knowledge in texts, lying at a level beneath the explicit assertional content. This knowledge consists of relationships implied to be possible in the world, or, under certain conditions, implied to be normal or commonplace in the world. The goal of the work reported is to derive such general world knowledge (initially, from Penn Treebank corpora) in two stages: first, we derive general "possibilistic" propositions from noun phrases and clauses; then we try to derive stronger generalizations, based on the nature and statistical distribution of the possibilistic claims obtained in the first phase. Here we report preliminary results of the first phase, which indicate the feasibility of our project, and its likely limitations.

1. INTRODUCTION

We think that there is a largely untapped source of general knowledge in texts, lying at a level beneath the explicit assertional content. This knowledge consists of relationships implied to be possible in the world, or, under certain conditions, implied to be normal or commonplace in the world. For instance, the sentence "*He entered the house through its open door*" suggests that it is possible for a person (or at least a male) to enter a house, that houses have doors, that doors can be open, etc. The goal of the present work is to derive such general world knowledge (initially, from Penn Treebank corpora) in two stages: first, we derive general "possibilistic" propositions from noun phrases and clauses; then we try to derive stronger generalizations, based on the nature and statistical distribution of the possibilistic claims obtained in the first phase.

A feature that our initiative shares with standard knowledge extraction work is its scalability and lack of dependence on deep semantic processing. However, it is distinctive in both its aims and its methodology. We are attempting to derive a broad range of general relationships from texts, rather than some predetermined specific kinds of facts; and we are using general phrase structure coupled with compositional interpretive rules to obtain general propositional information, rather than employing specialized extraction patterns targeted at specific relationships (e.g., [1, 2, 3, 6, 8, 14,19]) Our long-range goal is to use the derived knowledge as part of a KB supporting language understanding and commonsense reasoning.

We have reached a stage in our work where we are able to extract large numbers of general propositions from Treebank corpora, and can assess, in a very preliminary and informal way, the prospects for arriving at useful world knowledge by this method, and the limitations of the approach. In the following, we briefly sketch and assess our two-stage technique (with emphasis on the first, implemented stage).

2. EXTRACTING GENERAL "POSSIBILISTIC" PROPOSITIONS

The derivation of possibilistic propositions is based on a mechanism very similar to compositional semantic interpretation. The essential difference is that before combining the meanings of the immediate constituents of a phrase, we *abstract* those meanings, i.e., we simplify and generalize them; this involves stripping modifiers and inessential conjuncts, and generalizing individual terms (including named entities) to types. For example (using English glosses of the logical representations), abstraction of "*a long, dark corridor*" would yield "*a corridor*"; "*a small office at the end of a long dark corridor*" would yield "*an office*"; and "*Mrs. MacReady*" would yield "*a woman*". It is this process of abstraction, together with a weakening of the relations involved to a possibilistic form, that often yields general presumptions about the world, underlying the assertions made. At the same time, the fact that modifiers, conjuncts and specific meanings of definite descriptions are allowed to fall by the wayside as the interpretation of a sentence proceeds from lower-level to higher-level constituents greatly simplifies the interpretive process, in comparison with systems that attempt full understanding. Moreover, proposition extraction at lower and higher levels is relatively independent: interpretive failure at a lower level (e.g., for an "inessential" phrase, such as an adverbial) need not prevent proposition extraction from disjoint or higher-level phrases; and conversely, interpretive failure at a higher level leaves intact the propositions extracted from lower levels.

The choice of Treebank corpora as a basis for our initial knowledge extraction work is a natural one, since there exist robust, fairly accurate parsers trained on Treebank corpora (e.g., [7, 10, 9]), and these would eventually allow us to bootstrap our approach to very large unannotated corpora. The major challenge was that of providing reasonably reliable interpretive rules, at least for many of the clausal and other phrasal units that we viewed as potential sources of general possibilistic knowledge. In outline, our current algorithms for processing a bracketed (Treebank) sentence operate as follows:

*from *Human Language Technology 2002*, March 24-27, San Diego, CA, HLT 2002 Conference Notebook, pp. 84-87.

1. Preprocess the input tree (e.g., mark infinitives, passives, temporal noun phrases and prepositional phrases, categorize prepositional phrases, etc.)
2. Apply a set of ordered patterns to the tree recursively; these amount to phrase structure rules allowing for regular expressions (including negation) on the right-hand side
3. For each successfully matched subtree, abstract the interpretations of the essential constituents, and combine the abstracted interpretations in accord with the semantic rule linked to the pattern that matched the subtree
4. In processing the tree recursively, collect interpretations of phrases expected to provide general "possibilistic" propositions
5. Formulate possibilistic propositions from the collected phrasal interpretations, and output these along with simple English verbalizations.

The following is a simple example of a phrase structure pattern and associated semantic rule that could be used for a verb phrase in step 3:

(VP ((* ADVP (ADV) (PP)) ((V)) (NP) (S-INF) (*)))
(:f 1 (:p 2 3 4))

"Starred" constituents may occur 0 or more times. An example of a phrase from the Brown corpus matched (after preprocessing) by the phrase structure pattern is "*asked Kitti to join him*". The :f and :p in the semantic rule specify function and predicate application, and the digits are indices for the (abstracted) interpretations of the corresponding pattern constituents. (The possible rightmost, i.e., 5th, constituent or sequence of constituents is ignored by the semantic rule.) The semantic output for the phrase under consideration would be

(:p ASK[V] (:q DET FEMALE-INDIVIDUAL)
(:f KA (:p JOIN[V] (:q DET MALE-INDIVIDUAL)))),

whose gloss is "*ask a female individual to join a male individual*". (The :q indicates an unscoped quantifier, here the generic DET quantifier.)

An example of a complete and unedited output for a sentence from the Brown corpus is the following:

```
INPUT:

  ((S
    (S
      (NP (NNP Blanche) )
      (VP (VBD knew)
        (SBAR (\-NONE\- \0)
          (S
            (NP (NN something) )
            (AUX (MD must) )
            (VP (VB be)
              (VP (VBG causing)
                (NP
                  (NP (NNP Stanley) )
                  (POS \'s) (JJ new)
                  (\, \,)
                  (JJ strange) (NN behavior) )))))))
    (CC but)
    (S
      (NP (PRP she) )
      (ADVP (RB never) )
      (ADVP (RB once) )
      (VP (VBD connected)
        (NP (PRP it) )
        (PP (IN with)
          (NP (NNP Kitti) (NNP Walker) )))))
  (\. \.) )

OUTPUT:

WORDS OF THE INPUT:

(BLANCHE KNEW 0 SOMETHING MUST BE CAUSING STANLEY 'S NEW ,
 STRANGE BEHAVIOR BUT SHE NEVER ONCE CONNECTED IT WITH
 KITTI WALKER .)

OUTPUT (IN ENGLISH, FOLLOWED BY UNDERLYING LOGICAL FORMS):

 A FEMALE-INDIVIDUAL MAY KNOW A PROPOSITION.
 SOMETHING MAY CAUSE A BEHAVIOR.
 A MALE-INDIVIDUAL MAY HAVE A BEHAVIOR.
 A BEHAVIOR CAN BE NEW.
 A BEHAVIOR CAN BE STRANGE.
 A FEMALE-INDIVIDUAL MAY CONNECT A THING-REFERRED-TO WITH
   A FEMALE-INDIVIDUAL.

((:I (:Q DET FEMALE-INDIVIDUAL) KNOW[V] (:Q DET PROPOS))
 (:I (:F K SOMETHING[N]) CAUSE[V] (:Q THE BEHAVIOR[N]))
 (:I (:Q DET MALE-INDIVIDUAL) HAVE[V] (:Q DET BEHAVIOR[N]))
 (:I (:Q DET BEHAVIOR[N]) NEW[A])
 (:I (:Q DET BEHAVIOR[N]) STRANGE[A])
 (:I (:Q DET FEMALE-INDIVIDUAL) CONNECT[V]
     (:Q DET THING-REFERRED-TO)
     (:P WITH[P] (:Q DET FEMALE-INDIVIDUAL))))
```

In the logical forms, the :I indicates an infix formula, consisting of a subject, predicate, and possible additional arguments. Our logical forms are based on episodic logic [20], a natural logic that facilitates the transduction from language to logic and *vice versa*. As a second example, the following was obtained for a sentence of the Wallstreet Journal Corpus. (The input tree is omitted.)

```
(REP . RONNIE FLIPPO ( D . , ALA . ) , ONE OF THE MEMBERS
 OF THE DELEGATION , SAYS 0 HE WAS PARTICULARLY IMPRESSED
 *-1 BY MR . KRENZ 'S READY ADMISSION THAT EAST GERMANY
 NEEDED *-2 TO CHANGE .)

 AN ELECTED-REPRESENTATIVE MAY SAY A PROPOSITION.
 A DELEGATION MAY HAVE MEMBERS.
 A MALE-INDIVIDUAL MAY BE IMPRESS -ED BY AN ADMISSION.
 AN ADMISSION CAN BE READY.
 A COUNTRY MAY NEED TO CHANGE.

((:I (:Q DET ELECTED-REPRESENTATIVE) SAY[V] (:Q DET PROPOS))
 (:I (:Q DET DELEGATION[N]) HAVE[V]
     (:Q DET (:F PLUR MEMBER[N])))
 (:I (:Q DET MALE-INDIVIDUAL) (:F BE[PASV] IMPRESS[V])
     (:P BY[P] (:Q DET ADMISSION[N])))
 (:I (:Q DET ADMISSION[N]) READY[A])
 (:I (:Q DET COUNTRY) NEED[V] (:F KA CHANGE[V])))
```

These outputs (which are not atypical, quantitatively and qualitatively, for what we obtain for sentences from the Brown corpus or Wall Street Journal corpus) illustrates both the potential of our approach and some of the difficulties still to be overcome. Some of the propositions are arguably general knowledge that any human being is well-aware of, and which could be useful in a general NLU or commonsense reasoning system. For instance, the claims about who may say or know a proposition, about delegations having members, and about being impressed by an admission fall into this category.

On the other hand, a problem with our extracted propositions is that their content is often unclear or ambiguous (as in the case of "*An admission can be ready*", which intuitively allows for multiple senses of "*admission*" and "*ready*"), or they are true but arbitrary (e.g., the last of the extracted propositions for the Brown sentence certainly seems true enough, but completely arbitrary – not something likely to be useful for language processing or inference). Fortunately, outright falsehoods among the outputs are very rare. We are currently working out a judging scheme that will allow an empirical evaluation of the first-stage output, assessing the coherence, clarity/nonambiguity, nonarbitrariness, and truth of the output propositions, as determined subjectively by human judges (with sufficient consistency across judges). At the same time, we are considering various methods for improving the output, including use of WordNet [11] for type abstraction, better event-noun identification/classification, use of Treebank-3 information on argument roles, and most importantly, use of a coreference module to guess entity types for pronouns. We are also considering methods of filtering out unwanted output, e.g., discarding propositions as arbitrary if they are derived only once from a large corpus, even after certain abstraction operations.

3. THE SECOND STAGE: DERIVING STRONGER PROPOSITIONS

We have formulated several (as yet unimplemented) methods for strengthening some of the extracted propositions based on the types of symbols comprising them and on the occurrence statistics of related propositions. For example, we frequently extract certain simple variants of the first output proposition above, such as "A MALE-INDIVIDUAL MAY KNOW A PROPOSITION". From such sets of variants we expect to be able to infer that (by and large) *only* people or agencies know propositions. Similarly, preliminary examination of the Brown corpus indicates that we should be able to derive such generalities as (for transitive verb "*carry*"): if x carries y, then if y is a gun, then x is probably a person; if y is a bomb, then x is probably an airplane;[1] if y is a group of people, then x is probably a vehicle; etc. This sort of information is closely related to traditional *selectional preferences* (e.g., [5, 13, 16, 17, 21, 22]), but it is explicitly formulated as propositional knowledge by our methods and hence potentially usable not only for linguistic disambiguation but also for inference. Furthermore, extant methods for extracting selectional preferences from texts tend to "lose the connection" between the arguments of 2- or 3-place predicates (if they attempt to correlate multiple arguments at all). This is because of the myriad argument types that can occur (e.g., carriers found in Brown include persons, vehicles, [business] expansion, books, ants, positions, stories, roads, calculations, etc., and objects carried include [body]weight, passports, the economy, disclaimers, titles, responsibilities, traffic, rigging, etc.)

Our other planned methods of deriving stronger propositions depend on details of the propositions obtained in the first stage. For example, certain propositions involving possessive HAVE[V], such as "A FEMALE-INDIVIDUAL MAY HAVE AN ARM", derived from possessive NPs such as "*her arm*" can be directly and fairly safely strengthened to something like "A FEMALE-INDIVIDUAL GENERALLY HAS AN ARM" under certain conditions, for instance when the thing possessed is a part-type, or (less safely) when there is no possible prior coreferent for it.

4. CONCLUDING REMARKS

We have described a new method of obtaining general world knowledge from texts that is aimed at a broader range of facts than methods based on specialized extraction patterns, while still not depending on in-depth NLU.

It is of interest to compare the kinds of information obtainable by our method with the kinds obtained by specialized techniques. In essence, we obtain less specific information, but much more of it. For example, consider Berland and Charniak's techniques for extracting part-whole relations from text corpora [6]. Their extraction patterns relied primarily on genitives (such as "*the school's gymnasium*") and *of*-adjuncts (as in "*the basement of the building*"), and these are precisely the constructs from which we obtain propositions like the one mentioned above, that "*a female individual may have an arm*". However, note that at least in our first-stage processing we do not attempt to particularize HAVE to HAVE-AS-PART. Thus, from the phrase "*her sister*" we would derive the claim that "*a female individual may have a sister*", and as long as we interpret HAVE broadly, this is correct – whereas such an example is problematic if we are interested exclusively in part-whole relations. As indicated above, we intend in the second stage to appeal to occurrence statistics as well as certain kinds of lexical semantic information in deciding whether and how to strengthen first-stage propositions. At that point our output propositions may more nearly match those obtainable by specialized techniques; but to some extent we will continue to trade off specificity for breadth and generality.

Similar comments apply in relation to Girju and Moldovan's very interesting recent work on identifying sentences that express causal relationships [12]. Their targets were primarily sentences of form *NP1 causation-verb NP2*, e.g., "*Earthquakes cause tidal waves*", or "*The assassination led to World War I*" (though they also provide a quite comprehensive discussion of other relevant constructions). Again, our approach would succeed in extracting general propositions from such sentences (assuming we can parse them correctly), but would not, as it stands, explicitly distinguish a causal claims like AN ASSASSINATION MAY LEAD TO WAR from a non-causal one like A TRAIL MAY LEAD TO A LAKE. Girju and Moldovan make use of lexical information, such as that an assassination (unlike a trail) is a human action, to sort out causal from non-causal sentences. This is a prime concern for them, since their motivation for causal sentence identification is to enable text-based question-answering for causal questions such as "*What were the causes of World War I?*". For us, a more immediate concern is the extraction general presumptions, in an explicit logical form, from specific (factual or fictitious) texts, with maximally broad coverage of the relationships encounterd. However, we would expect to use Girju and Moldovan's type of lexical semantic information in deploying our extracted propositions for inference.

One extant system for more general knowledge acquisition from text is MindNet [18]. However, the goal of this system is not the inference of general propositions from specific texts, but rather the direct interpretation of the contents of MRDs such as the *American Heritage Dictionary, 3rd Edition*. As such, it is reported to achieve full coverage, though the available descriptions of its design, capabilities and limitations are somewhat sketchy. In any case, it would certainly be desirable to merge lexicon-derived knowledge with knowledge inferred from texts, since the types of information obtainable by these methods are quite distinct. (A glance at our sample ouput indicates that few of our derived propositions are of the sort found in lexicons.)

As we have indicated, there are some similarities between our derived propositions and the information implicit in selectional preferences. Efforts to extract predicate-argument structure are also related (e.g., Abney's method based on a cascaded finite-state parser [4][2]). However, the attempt to formulate and evaluate the derived information as general propositions about the world is to our knowledge novel.

The quest for general world knowledge also brings to mind CYC [15], but as in the case of MindNet we regard that work as complementary rather than an alternative. Little of the knowledge we are extracting can be found in CYC; and little of the knowledge in CYC is obtainable by our methods. An impediment to merging the two kinds of knowledge, however, is that CYC's invented (and rather heterogeneous) ontology is not easily brought into alignment with linguistically derived knowledge.

[1] For more extensive or recent texts, x may with some likelihood be a person.

[2] Our experimentation on some Brown corpus sentences suggests that this approach yields very sparse information – for many sentences, no more than a subject-verb relation, because of the parser's very conservative phrase attachment policy.

Acknowledgements

This work was supported by the National Science Foundation under grant IIS-0082928, and greatly benefited from discussions with David Ahn, Greg Carlson, Aaron Kaplan, and Henry Kyburg.

5. REFERENCES

[1] *Proc. of the 5th Message Understanding Conference (MUC-5)*. Morgan Kaufmann, Los Altos, CA, 1993.

[2] *Proc. of the 6th Message Understanding Conference (MUC-6)*. Morgan Kaufmann, Los Altos, CA, 1995.

[3] *Proc. of the 7th Message Understanding Conference (MUC-7)*. Morgan Kaufmann, Los Altos, CA, April 29 – May 1, Virginia 1998.

[4] S. Abney. Partial parsing via finite-state cascades. In *Proceedings of the ESSLLI '96 Robust Parsing Workshop*. 1996.

[5] E. Agirre and D. Martinez. Learning class-to-class selectional preferences. In *Proc. of the 5th Workshop on Computational Language Learning (CoNLL-2001)*, Toulouse, France, July 6-7, 2001.

[6] M. Berland and E. Charniak. Finding parts in very large corpora. In *Proc. of the 37th Ann. Meet. of the Assoc. for Computational Linguistics (ACL-99)*, Univ. of Maryland, June 22 - 27, 1999.

[7] E. Charniak. A maximum-entropy-inspired parser. In *Proc. of the 1st Meet. of the North Am. Chapt. of the Assoc. for Computational Linguistics (NACL-2000)*, pages 132–139, Seattle, WA, April 29 - May 4, 2000.

[8] S. Clark and D. Weir. An iterative approach to estimating frequencies over a semantic hierarchy. In *Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999. Also available at http://www.cogs.susx.ac.uk/users/davidw/research/papers.html.

[9] M. J. Collins. Three generative, lexicalised models for statistical parsing. In *Proc. of the 35th Ann. Meet. of the Assoc. for Computational Linguistics (ACL-97) and 8th Conf. of the Eur. Chapter of the Assoc. for Computational Linguistics (EACL-97)*, pages 16–23, 1997.

[10] M. J. Collins. Discriminative reranking for natural language parsing. In *Proc. of the 7th Int. Conf. on Machine Learning (ICML-2000)*, 2000.

[11] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[12] R. Girju and D. I. Moldovan. Text mining for causal relations. In *FLAIRS 2002*, 2002.

[13] R. Grishman and J. Sterling. Acquisition of selectional patterns. In *Proc. of COLING-92*, pages 658–664, Nantes, France, 1992.

[14] M. A. Hearst. Automated discovery of WordNet relations. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 131–153? MIT Press, 1998.

[15] D. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Comm. of the ACM*, 38(11):33–38, 1995.

[16] P. Resnik. A class-based approach to lexical discovery. In *Proc. of the 30th Ann. Meet. of the Assoc. for Computational Linguistics (ACL-92)*, pages 327–329, Newark, DE, 1992.

[17] P. Resnik. Semantic classes and syntactic ambiguity. In *Proc. of ARPA Workshop on Human Language Technology*, Plainsboro, NJ, 1993.

[18] S. D. Richardson, W. B. Dolan, and L. Vanderwende. MindNet: acquiring and structuring semantic information from text. In *Proc. of the 36th Ann. Meet. of the Assoc. for Computational Linguistics and the 17th Int. Conf. on Computational Linguistics (COLING-ACL'98), vol. II*, volume 2, pages 1098–1102, Montreal, Canada, Aug. 10-14, 1998.

[19] E. Riloff and R. Jones. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proc. of the 16th Nat. Conf. on Artificial Intelligence (AAAI-99)*, 1999.

[20] L. K. Schubert and C. H. Hwang. Episodic Logic meets Little Red Riding Hood: A comprehensive, natural representation for language understanding. In L. Iwanska and S. Shapiro, editors, *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*, pages 111–174. MIT/AAAI Press, 2000.

[21] U. Zernik. Closed yesterday and closed minds: Asking the right questions of the corpus to distinguish thematic from sentential relations. In *Proc. of COLING-92*, pages 1304–1311, Nantes, France, Aug. 23-28, 1992.

[22] U. Zernik and P. Jacobs. Tagging for learning: Collecting thematic relations from corpus. In *Proc. of the 13th Int. Conf. on Computational Linguistics (COLING-90)*, pages 34–39, Helsinki, 1990.