Extracting Word Relationships from Unstructured Data (Learning Human Activities from General Websites)

Anirudha S. Bhat [a], **Krithika K. Iyer** [b], **and Rahul Venkatraj** [c]

Abstract

One of the biggest challenges of instructing robots in natural language, is the conversion of goals into executable instructions [1]. A core problem in this area is to train the robot to identify the specific actions (verbs) that help fulfil the goal using the objects (nouns) at its disposal. This project presents a new algorithm, which we call "Extracto", to extract noun-verb-noun relationships from unstructured text. Extracto is built on the Coupled Pattern Learner framework [2,3] presented by the creators of the 'Never Ending Language Learner' (NELL) project. Extracto gives a 70% precision in identifying suitable action verbs, given two nouns. We have extended this to create the "Extracto-Categorize" algorithm that extracts actions for the categories of the respective nouns instead of the nouns themselves. This increases extraction precision to 82%. Since this is a first of its kind method created purely for this purpose (noun-verb-noun extraction), we have benchmarked it against related algorithms that extract semantic relationships from unstructured text. Existing algorithms reach precision levels of 70 - 91% [4,5,6,7], which is in line with Extracto's performance.

1 Introduction

Robots are advancing rapidly in their behavioural functionality allowing them to perform sophisticated tasks. However, their ability to take Natural Language instructions is still in its infancy. Parsing, Semantic Intrepretation and Dialogue Management are typically performed only on a limited set of primitives, thus limiting the set of instructions that could be given to a robot. This limits a robot's applicability in unconstrained natural environments (like households and offices) [8].

In this project, we are only addressing the problem of semantic interpretation of human instructions. Specifically, our Extracto algorithm provides a method to extract potential actions (verbs) that could be performed given two household objects (nouns). For example, given the nouns "Coffee" and "Cup", Extracto identifies the action (verb) "pour" indicating that 'coffee should be poured in a cup', and not 'stored' or 'roasted'. A human instruction "I want coffee" or "Get me a cup of coffee" is a goal for the robot, but does not specifically instruct the robot what to do with the cup and the coffee. The Extracto method helps address this particular problem. In addition, given an action (verb), Extracto identifies the most suitable objects (nouns) to perform the task. For example, given an action (verb) "introduce", Extracto identifies a series of suitable noun pairs, one of which is "friend" and "host" which means 'a friend is to be introduced by a host'.

The rest of our report is structured as follows: In Section 2, we take you through a short summary of past work done in this area. Next, in Section 3 and 4, we explain our data sources and processing steps. Section 5 gives a detailed explanation about Extracto and its extension to 'Extracto-Categorize'. This is followed by a summary of our key results in Section 6. To conclude we present our ideas on future potential in Section 7, our acknowledgements in Section 8 and references in Section 9.

2 Related Work

Fundamentally, there are three approaches to extracting semantic word relationships: (1) supervised, (2) unsupervised and (3) bootstrap learning starting with very small seed instances. A supervised approach would involve labelling entities in a text corpus, and then training classifiers to capture relations between pairs of entities in a sentence or a combination of sentences [9,10,11]. This is expensive and the resulting classification algorithm would be biased towards the particular corpus. Unsupervised approaches extract abundantly many strings (for subsequent relation-extraction) from large amounts of text, but it would be difficult to map extracted relations to a particular knowledge base [12,13]. The third approach starts with a small set of seed instances, which generate more instances which in turn generate even more instances iteratively [3,4,14,15,16]. This approach helps us overcome the shortcomings of the first two approaches, and hence we chose this.

0 [a] *asbhat@stanford.edu*
0 [b] *kiyer@stanford.edu*
0 [c] *vrahul@stanford.edu*

There have been several past attempts to extract semantic relationships using the third approach (starting with a seed set), but none have been specifically aimed at extracting noun-verb-noun relationships which we are targeting. Some of the most relevant studies are indicated here. Carlson et al. [2,3] present a framework for semi-supervised learning in a similar context (NELL). Specifically, their Coupled Pattern Learner (CPL) methodology lays out the broad framework of how one can extract word-relationships given an ontology and a text corpus. The CPL's precision ranges from 70% to 100% for a variety of relationships (For example, 'X is a building', 'Y is a conference', 'Company P produces product Q' etc.), averaging a precision of 89%. Pantel et al. present an algorithm [4] that is designed to harvest "Is-A" (protein::biopolymer), "Part-Of" (oxygen::air) relations along with succession (Ford::Nixon), reaction (boron::flourine) and production (kidney::kidney stones) pairs. They achieve a precision of 80% on news-articles-text and 91% precision on a chemistry textbook. Similar to our Extracto algorithm, the three core steps of their algorithm are Pattern Induction, Ranking/Selection and Instance Extraction.

3 Data Sources

The set of seed instances (30 in number) were constructed manually from our team's understanding of regular kitchen and home behaviour of humans. This comprises noun-verb-noun relations like 'cup-drink-coffee'.

Table 1 Examples of original seed triads

Action (verb)	Noun	Noun
drink	coffee	cup
seat	lady	table
eat	spoon	plate
pass	plate	knife
open	gentleman	door

Unstructured data about relevant nouns (say, 'coffee') were obtained by crawling the respective pages on the domains: en.wikipedia.org, wikihow.com, simple.wikipedia.org, thefreedictionary.com, yourdictionary.com. The crawlers (spider programs) were built using the Python library 'Scrapy'. In addition, full books and novels of relevant topics from Project Gutenberg were taken. Some examples include "The Ladies' Book of Etiquette, and Manual of Politeness by Florence Hartley", "Etiquette by Emily Post" and "Manners and Rules of Good Society Or Solecisms to be Avoided". In 'Extracto-Categorize', we use categories of the respective nouns, instead of the nouns themselves. For example, categories of 'coffee' would include 'beverage' and 'liquid' among others (as shown in Figure 1). These are sourced recursively from the Hypernym relations given by Princeton WordNet [17].

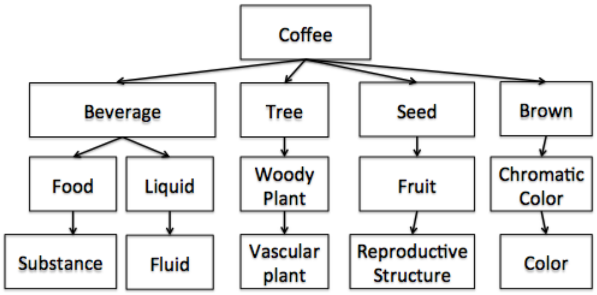
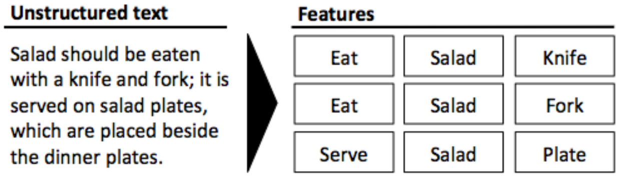


Fig. 1 Coffee categories (Hypernyms) from WordNet[17]

4 Features and Pre Processing

The noun-verb-noun triads are the equivalent of "features" for this method. These are extracted using the Syntactic Tree output from the Stanford Parser [18] and the Parts-of-Speech tagger from the Natural Language Toolkit (NLTK) Python libraries. Also, the NLTK Lemmatizer [19] (looking up Princeton WordNet) is used to lemmatise nouns and get to their tense-less root forms. For example, 'poured', 'pouring' and 'pours' are all converted to 'pour'.



Unstructured text

Salad should be eaten with a knife and fork; it is served on salad plates, which are placed beside the dinner plates.

Features

Eat	Salad	Knife
Eat	Salad	Fork
Serve	Salad	Plate

Fig. 2 Extracting Features from Unstructured Text

5 Models

As a quick proof of concept, we did a traditional Bag-Of-Words exercise. We split unstructured data into buckets (roughly, paragraphs), and estimated the probability of a verb (say, "pour") occurring given two nouns (say, "cup"and "coffee"). The verb relation is considered valid if the probability of the specific verb occurring, given the two nouns is greater than 0.5. After we observed promising results using the Bag Of Words approach, we implemented and refined our Extracto algorithm.

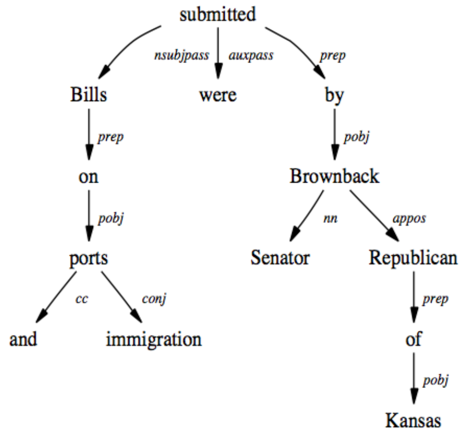


Fig. 3 A sample syntactic tree [18]

At the very core, the Extracto algorithm identifies actions (verbs) that fit with the seed nouns from unstructured text, and then finds some more new nouns that fit with the newly found actions (verbs). The noun-verb-noun relations are ranked and then best few are added to the seed set as inputs to the next iteration. This way, Extracto predicts more and more noun-verb-noun triads iteratively.

Table 2 Extracto Method - pseudocode

Input: Seed set of relationships, unstructured text corpus
Output: Predicted set of Noun-Verb-Noun relationships

```
Repeat full sequence till no new triads generated
  For each Noun-Noun pair in the Seed set
    Identify verbs suited to respective Noun-Noun pair
    Extract new Noun-Noun pairs with newly obtained verbs
    Rank the set of triads obtained
    Update seed set with highest ranked triads
    Promote the highest ranked triads to seed set
```

A major step of Extracto is the identification of semantically related nouns and verbs from unstructured data. We parse each sentence using the Stanford Parser [18] to obtain a Syntactic Tree relationship similar to Figure 1. For example, 'Noun Subject' (nsubj) and 'Objective' (dobj) relations are very clear associations between verbs and nouns. In the second step of identifying more nouns that fit the extracted action (verb), we use the NTLK POS Tagger to capture the full list of nouns in the respective sentences. Since the most relevant of these outputs are taken and added as feed to the next iteration, every iteration produces more relationships with increased accuracy, as shown in the Figure 4. Further (and intuitively), the precision of results predicted increases when larger seed sets are used (Figure 5).

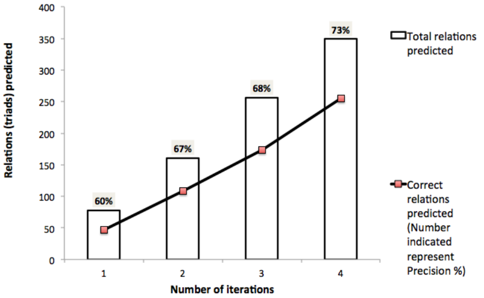


Fig. 4 Increasing relations and precision with every iteration

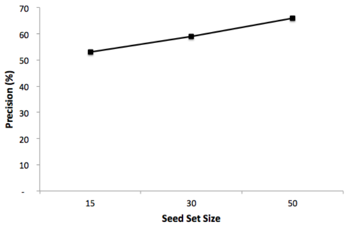


Fig. 5 Increasing precision with larger seed set sizes

In order to improve the precision of extracted relationships, we have implemented a novel extension to the Extracto method. We call it 'Extracto-Categorize'. We could substantially improve relationship capture by looking for the noun's category instead of the noun itself in the unstructured text corpus. For example, we are looking for the relevant verbs (like "pour") when we try to connect "coffee" and "cup". Instead of searching for 'coffee' and 'cup', we are more likely to find a relationship (in any unstructured text) if we search for 'any beverage'and 'any container'. So, even if the text contains sentences which leads to links like 'milk-pour-glass', 'beer-pour-mug', 'wine-pour-chalice' or 'tea-pour-cup', the "pour" action (verb) will get captured for 'coffee' and 'cup'.

6 Results and Discussion

The success metric we have chosen is Precision, which we define as the number of correct relationships extracted, among all the relationships finally produced by Extracto. The correct relationships were manually identified and accepted as correct if an internet search of the three words together resulted in the intended action being the first result.

$$Precision = \frac{Number\, of\, relevant\, results\, extracted}{Total\, number\, of\, results\, extracted} \quad (1)$$

Extracto led to the following results: Precision of predicting verbs given nouns was 71%, and precision of predicting nouns given verbs was 80%. Some good example relations extracted are shown in Table 3, and some wrong ones extracted are shown in Table 4.

We extended the Extracto algorithm and implemented the Extracto-Categorize method to improve the precision of predicting verbs given nouns. With this extension, the precision improved from 71% to 82%. Some of the new relations obtained after this extension are shown in tables 5 (useful results) and 6 (incorrect results).

Table 3 Useful relationships predicted by Extracto

Action (verb)	Noun	Noun
put	napkin	knees
pour	tea	cup
pour	coffee	cup
introduce	hostess	person
offer	fees	servant
offer	lady	seat
cut	fruit	knife
eat	food	fork
drink	men	soup
use	bread	butter
put	sugar	spoon

Table 4 Incorrect relationships predicted by Extracto

Action (verb)	Noun	Noun
eat	meal	state
prefer	part	piece
blow	men	soup
carve	fork	food

Table 5 Useful relationships predicted by Extracto-Categorize

Action (verb)	Noun	Noun
pass	tea	cup
pass	coffee	cup
escort	lady	table
use	fork	hand

Table 6 Incorrect relationships predicted by Extracto-Categorize

Action (verb)	Noun	Noun
help	article	table
prefer	name	piece
be	dish	table

7 Conclusion and Potential for Future Work

We believe our 'Extracto' and 'Extracto-Categorize' algorithms provide a quick and robust mechanism to predict word relationships from unstructured text data. This is a key step in taking natural language instructions to robots, to the next level. Although no other algorithm addresses the requirement at this level of specificity, our 82% precision is found to be in line with other broader and related algorithms that extract word relationships from unstructured data.

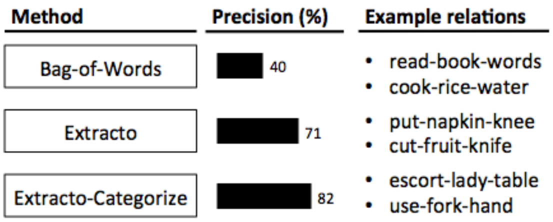


Fig. 6 Summary of results : Predicting Actions (verbs)

The immediate next step on this would be the representation of all predicted relationships in a network / graph of relationships, that would be easily understood by a robot. A cost function estimate would help determine the most suitable action (verb) given a host of potential options. In the future, Extracto could be extended to obtain relationships for other areas outside the household as well (extensions include hospital assistance, sports training, office spaces and laboratories). For this to be successful, we will need to create a mechanism for automatically choosing the websites to crawl, and the books to read.

8 Acknowledgement

We would like to thank Professor Ashutosh Saxena and Dipendra Kumar Misra for their guidance and support.

9 References

[1] Stenmark, M., & Nugues, P. (2013, October). Natural language programming of industrial robots. In Robotics (ISR), 2013 44th International Symposium on (pp. 1-5). IEEE.

[2] Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E. R., & Mitchell, T. M. (2010, July). Toward an Architecture for Never-Ending Language Learning. In AAAI (Vol. 5, p. 3).

[3] Carlson, A., Betteridge, J., Wang, R. C., Hruschka Jr, E. R., & Mitchell, T. M. (2010, February). Coupled semi-supervised learning for information extraction. In Proceedings of the third ACM international conference on Web search and data mining (pp. 101-110). ACM.

[4] Pantel, P., & Pennacchiotti, M. (2006, July). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (pp. 113-120). Association for Computational Linguistics.

[5] Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009, August). Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2 (pp. 1003-1011). Association for Computational Linguistics.

[6] Lee, C. S., Kao, Y. F., Kuo, Y. H., & Wang, M. H. (2007). Automated ontology construction for unstructured text documents. Data & Knowledge Engineering, 60(3), 547-566.

[7] Hawizy, L., Jessop, D. M., Adams, N., & Murray-Rust, P. (2011). ChemicalTagger: A tool for semantic text-mining in chemistry. Journal of cheminformatics, 3(1), 17.

[8] Dzifcak, J., Scheutz, M., Baral, C., & Schermerhorn, P. (2009, May). What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In Robotics and Automation, 2009. ICRA'09. IEEE International Conference on (pp. 4163-4168). IEEE.

[9] ZHOU12, G., Zhang, M., Ji, D. H., & Zhu, Q. (2007). Tree kernel-based relation extraction with context-sensitive structured parse tree information. EMNLP-CoNLL 2007, 728.

[10] GuoDong, Z., Jian, S., Jie, Z., & Min, Z. (2005, June). Exploring various knowledge in relation extraction. In Proceedings of the 43rd annual meeting on association for computational linguistics (pp. 427-434). Association for Computational Linguistics.

[11] Surdeanu, M., & Ciaramita, M. (2007, March). Robust information extraction with perceptrons. In Proceedings of the NIST 2007 Automatic Content Extraction Workshop (ACE07).

[12] Shinyama, Y., & Sekine, S. (2006, June). Preemptive information extraction using unrestricted relation discovery. In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (pp. 304-311). Association for Computational Linguistics.

[13] Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007, January). Open information extraction for the web. In IJCAI (Vol. 7, pp. 2670-2676).

[14] Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., ... & Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. Artificial Intelligence, 165(1), 91-134.

[15] Bunescu, R. C., & Mooney, R. (2007, June). Learning to extract relations from the web using minimal supervision. In Annual meeting-association for Computational Linguistics (Vol. 45, No. 1, p. 576).

[16] Rozenfeld, B., & Feldman, R. (2008). Self-supervised relation extraction from the Web. Knowledge and Information Systems, 17(1), 17-33.

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

[18] Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In LREC 2006.

[19] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. OReilly Media Inc.