

Learning Incident Causes

Amandine Orecchioni, Nirmalie Wiratunga, Stewart Massie, Sutanu Chakraborti, and
Rahman Mukras

School of Computing, RGU,
Aberdeen AB25 1HG, Scotland, UK
{a.o|nw|sm|sc|ram}@comp.rgu.ac.uk

Abstract. This paper analyses the Air Investigation Reports corpus available from the Transportation Safety Board of Canada. It proposes a machine learning approach to domain specific sentence classification as a first step towards report matching. A structured case representation at the sentence level is created using Link Grammar Parser and WordNet and compared to a bag-of-words representation. Preliminary results are favourable towards the structured representation.

1 Introduction

The Air Investigation incident reports¹ form a useful source of referential knowledge for future incident investigations motivating research into textual case comparison mechanisms. Retrieving similar reports is challenging because investigators may typically want to retrieve reports based on multiple factors such as weather conditions, aircraft type, geographical location or cause of the incident. Although each of these factors can be manually extracted from sentences, the number of incidents and sentences makes this a daunting task. In this paper we propose to automate sentence classification according to one of these factors. We see this task as a first step towards automatically identifying predefined factors for case comparison. Here we adopt a binary classification where a sentence is labelled as either containing causal information or factual information. For instance, “*The aircraft began to accumulate ice at a rate that exceeded capabilities of the ice-protection equipment*” would be labelled as a *cause* while “*The aircraft departed at 0537 central daylight time*” would be labelled as a *fact*.

A manually labelled corpus of classified sentences is used to test whether a sentence level structured representation helps capturing underlying meaning. The representation decomposes a sentence into its constituent parts - subject, object, verb and modifiers. Our experiments show that a classifier is better able to differentiate between sentence classes when the structured representation is used rather than a bag-of-words (BOW). This paper also discusses the utility of feature vector generalisation using background knowledge from WordNet [3]. Word similarities obtained by analysing the organisation of words within WordNet helps reduce sparse representations so that sentences containing different yet semantically similar words are correctly considered similar.

We introduce the terminology used to describe our case representation in Section 2. The creation of our experimental dataset by converting free text into a structured rep-

¹ Transportation Safety Board of Canada: <http://www.tsb.gc.ca/en/reports/>

representation is discussed in Section 3. An evaluation of our approach is reported in Section 4 followed by conclusions in Section 5.

2 Case Representation

The popular BOW representation technique ignores word order during text comparison. For example, “*The pilot rescued the passenger*” and “*The passenger rescued the pilot*” differ in meanings but have the same BOW representation. The limitations of BOW and the need for a better text case representation has previously been identified. IE and NLP techniques were used to generalise proper nouns into their roles to facilitate case comparison [1]. In this paper, we are not concerned by the presence of proper nouns but more by the semantic structure of sentences.

We propose a Structured Sentence Representation (SSR) to map a sentence into 4 attributes: subject, verb, object and modifiers. A classified sentence is a pair (x, y) where $x = \{\vec{S}, \vec{V}, \vec{O}, \vec{M}\}$ is a set of 4 attributes and y denotes the class label. Here $\vec{S} = (f_1, \dots, f_{S_n})$ is a binary valued feature vector corresponding to the presence or absence of keywords in the subject part of a sentence. \vec{V} , \vec{O} and \vec{M} are similarly defined for the verb, object and modifier parts of a sentence. This representation allows sentence comparison at the constituent level (e.g. subject with subject, verb with verb).

Table 2 shows a fictitious casebase of four classified sentences represented using SSR. The meaning of s_1 and s_2 is preserved by comparing *The pilot* with *The passenger* as subject, *rescued* with *rescued* as verb and *the passenger* with *the pilot* as object. SSR would therefore differentiate these sentences by their subjects and objects whereas BOW would represent them as identical.

Sentences	SSR Representation														Class y					
	\vec{S}					\vec{O}					\vec{V}			\vec{M}						
	pilot	passenger	helicopter	runway	edge lights	rescued	was	carrying	out	were	of	order	passenger	pilot		lake	water	sampling	aircraft	
s_1 ="The pilot rescued the passenger from the aircraft"	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	fact
s_2 ="The passenger rescued the pilot"	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	fact
s_3 ="The runway edge lights were out of order"	0	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	cause
s_4 ="The helicopter was carrying out lake water sampling"	0	0	0	1	1	0	1	1	1	0	0	0	0	1	1	1	0	0	0	fact

Table 1. SSR representation of a fictitious casebase of four cases

3 Dataset creation

A dataset was created by extracting single sentences from the “*Summary*” section and the “*Findings as of Causes*” section of the 12 reports dating from 2005. Only 2 sections were selected because each sentence was then manually tagged as *cause* or *fact*. These sections were assumed to provide a representative set of sentences for each class. The final dataset contains 240 instances, including 54 causes and 186 facts. It represents a vocabulary of 719 features for the BOW representation, 129 for the subjects, 190 for the verbs, 132 for the objects and 333 for the modifiers. Once labelled, sentences are decomposed into their constituent parts using the Link Grammar Parser [6].

3.1 Sentence Decomposition with the Link Grammar Parser

The Link Grammar Parser is a dictionary-based parser for the English language where each word is associated with one or more links, together with rules detailing how neighbouring words in the sentence might be linked. When the parser is presented with a sentence, it assigns a syntactic structure, which consists of a set of labelled links connecting pairs of words (see Figure 1). For instance the words *helicopter* and *was* are linked by the Ss link in Figure 1, denoting a verb-subject relationship. The parser also identifies the part of speech for each word in the sentence and produces a constituent tree including noun phrases (NP) and verb phrases (VP).

We use the heuristics set out in [4] to extract the main verb of a sentence, its subject and its object. For instance, an obvious rule to identify the subject is to find a noun which makes an S link with a verb (e.g. *helicopter*–*was*). We have extended the heuristics so that subject-object-verb relationship extraction is possible at both the word and phrase level. In order to do this, we use the phrase information in the constituent tree. Consider the example in Figure 1, the object NP “*lake water sampling*” is identified by extracting the object word “*sampling*” and identifying the smallest NP containing this word. Similarly, we select the VP (excluding any NPs) corresponding to the identified verb. In Figure 1, the rules identify the main verb as “*carrying*” and the extension will extract “*was carrying out*” as the VP.

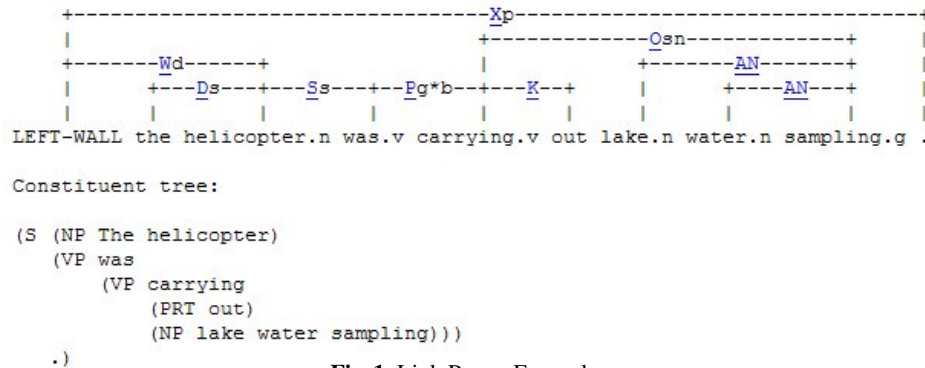


Fig. 1. Link Parser Example

3.2 Expansion using WordNet

SSR representation is better able to capture meaning as each sentence part is treated independently. However, it increases sparseness since each constituent part contains fewer words than the entire sentence. This problem is further aggravated by the fact that some sentences may not have a subject part or an object part. For those sentences that do have subjects and/or objects, small differences in vocabulary can contribute adversely to the similarity computation. Problems due to variability in vocabulary have previously been addressed using text generalisation [7].

We propose to generalise each SSR vector of a sentence using WordNet [3]. Values of absent features are boosted based on their semantic relatedness to present features. This is achieved by aggregating the current value of a feature with its semantic relat-

edness to present features using Mycin’s approach [2]. The algorithm to generalise a binary vector $\vec{v} = (f_1, \dots, f_n)$ appears in Figure 2.

```

for i=0 to n
  if  $f_i=0$  /* Absent feature */
    then for j=0 to n
      if  $f_j=1$  /* Present feature */
        then if  $SemRel(f_i, f_j) \geq 0.8$ 
          then  $f_i = Agg[(f_i, SemRel(f_i, f_j))]$ 

```

Fig. 2. Generalisation algorithm

The semantic relatedness of two words is typically ascertained by counting the nodes distancing words and the depth of their common ancestor node. Studies in [5] report good correlation between human judgement and the Wu & Palmer measure [8], where relatedness between f_i and f_j is defined as follows:

$$SemRel(f_i, f_j) = \frac{2D_{lca}}{D_{f_i} + D_{f_j}} \quad (1)$$

D denotes the depth of a particular node in relation to the root and LCA represents the least common ancestor between 2 words, which is their first common node. The closer the words are to their LCA, the more semantically related they are.

4 Evaluation & Results

We evaluated our approach on the binary sentence classification task using k NN with leave-one-out testing. BOW and SSR representations were created with and without feature selection using information gain. Case retrieval with SSR involved separate similarity computations for each constituent part. A combined similarity score is then obtained by allocating to each constituent part a weight discovered by a genetic algorithm. SSR’s constituent parts were also separately evaluated to establish the influence of subject, object, verb and modifier on classification accuracy.

	BOW	<i>Subject</i>	<i>Object</i>	<i>Verb</i>	<i>Modifier</i>	SSR
<i>Base</i>	0.8	0.8	0.79	0.74	0.78	0.81
<i>WordNet</i>	0.78	0.8	0.76	0.75	0.8	0.83
<i>WordNet + Feature Selection</i>	0.8	0.83	0.78	0.8	0.83	0.84

Table 2. Overall accuracy

Overall accuracy for the cause identification task varies between 74% and 84%. As expected SSR, both generalised and subjected to feature selection, results in highest accuracy. We note that with 90% vocabulary coverage in this domain, WordNet is a good source for text generalisation. The lowest accuracy results were obtained with BOW when using just the verb part of sentences. WordNet-based generalisation has a favourable impact on all but the BOW and the object representations. Closer examination of the object results indicated that generalisation blurs the vocabulary distinction between causes and facts. This suggests that the object vocabulary may not be as

discriminatory of the cause and fact classes compared with the verb, subject or modifier vocabulary. The poor performance with the object representation is not unexpected since 50% of the sentences in the casebase have no object part.

We used precision and recall to evaluate the performance of the system. Precision is generally low, with a maximum of 43% with BOWGEN. Recall of 89% was achieved when classification is based only on the modifier part of sentences with generalisation and feature selection. However, considering the biased nature of the casebase, precision and recall independently failed to present a conclusive result, except when combined by the F-measure (harmonic mean), where a maximum of 51% was achieved by generalised SSR with feature selection.

5 Conclusions and Future Work

In this paper, we have investigated the benefits of using a structured representation over BOW for a sentence classification task. Our preliminary results show that SSR improves accuracy over BOW. Overall accuracy benefits from feature generalisation and feature selection. However moderate benefits are seen when constituent parts are treated independently, suggesting a loss of contextual knowledge which is preserved only when combined in SSR. More extensive experiments on a balanced dataset would have to be carried out as future work in order to achieve statistically significant results. Experimenting on a different corpus would also evaluate how applicable this technique is to other domains.

The manual labelling of sentences was difficult because it was often unclear whether a sentence contained causal or factual information, or both. In future work, we would like to investigate the usefulness of the SSR approach to identify other case comparison factors such as aircraft type and weather conditions, where manual labelling is less demanding. Furthermore, since multiple factors can be contained in single sentences, an investigation at the phrase level will be useful.

References

1. Brüninghaus, S. and Ashley, K. D. The role of Information Extraction for Textual CBR. In *ICCB* pp.74-89, 2001
2. Davis, R., Buchanan B., and Shortliffe, E. Production Rules as a Representation for a Knowledge-Based Consultation Program. *Artificial Intelligence* 8 pp.15-45, 1977
3. Fellbaum, C., editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998
4. Harsha V. Madhyastha, N. Balakrishnan, and K. R. Ramakrishnan. Event Information Extraction Using Link Grammar. In *RIDE-MLIM* pp.16-22, 2003
5. van der Plas, L. and Bouma, G. Syntactic contexts for finding semantically related words. In *CLIN* pp.173-186, 2004
6. Sleator, D. and Temperley, D. Parsing english with a link grammar. In *IWPT* pp.277-292, 1993
7. Wiratunga, N., Koychev, I., and Massie, S. Feature Selection and Generalisation for Retrieval of Textual Cases. In *ECCBR* pp.806-820, 2004
8. Wu, Z. and Palmer, M. Verb semantics and lexical selection. In *ACL* pp.133-138, 1994.