Automated Extraction of Concept Features and Relations

Colin Kelly

1. Introduction

How would you describe an aeroplane? How would someone else? What distinguishes it as an object?

Our aim is to produce a system which, given a noun concept (e.g. "aeroplane") is able to output a list of real world knowledge (basic facts about that concept) preferably of the type humans would use to describe that concept.

For example, we might want to output "aeroplanes have wings", "aeroplanes are fast", "aeroplanes are used for transportation", and so on.

We hope this research will prove useful in a wide variety of applications, including the possibility of building a relational semantic network as well as leading to a better understanding of how the brain stores concepts and their relationships.

2. Features and Relations

Typically, humans will describe noun concepts in a wide variety of ways. For example, when asked to describe *an aeroplane* they will say such things as:

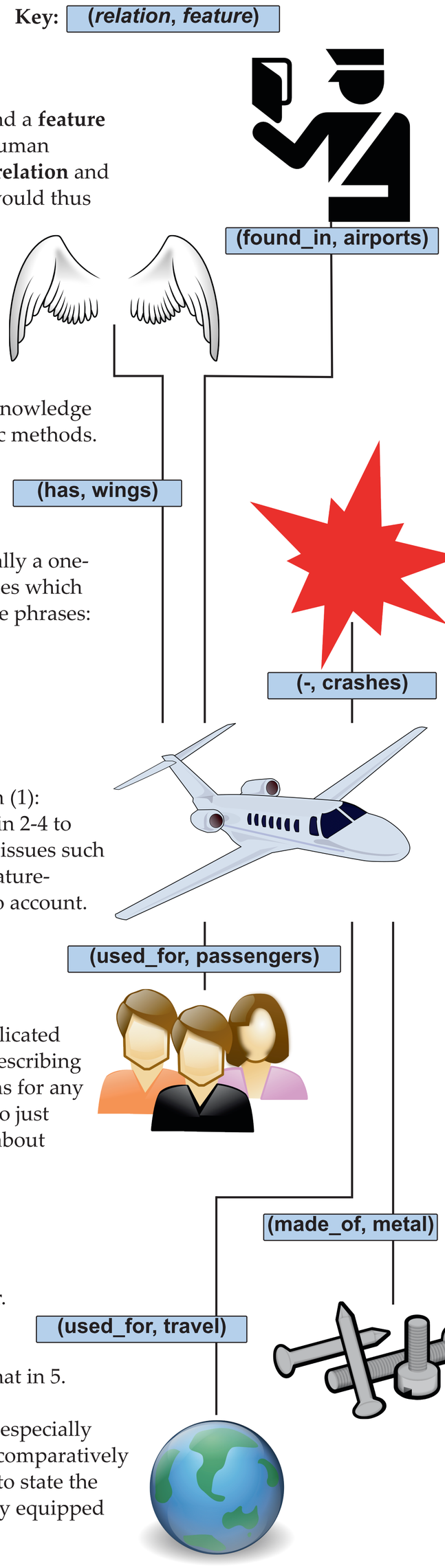
An aeroplane is used for transportation.

From this statement we may derive a **relation** (***used_for***), and a **feature** (***transportation***). It is possible to translate the majority of human statements about non-abstract nouns into a list of **concept**, **relation** and **feature** triples. Some typical statements about aeroplanes would thus translate to:

(**concept, relation, feature**)
(***aeroplane, used_for, transportation***)
(***aeroplane, has, wings***)
(***aeroplane, made_of, metal***)

The aim of our project is to generate such true, real-world knowledge triples using large text corpora and computational linguistic methods.

Figure 1:
Describing an aeroplane



2a. Features

This is the easier of the two parts to derive: a feature is usually a one-word noun or adjective. However there still some basic issues which must be dealt with. For example, if our corpus contained the phrases:

1. An aeroplane is fast.
2. An airplane is fast. Aircraft are fast.
3. An aeroplane is not slow. / An aeroplane is speedy.
4. Aeroplanes travel at high speed.

We would presumably want to produce only the triple from (1): (***aeroplane***, ***is***, ***fast***). But we would also want the sentences in 2-4 to reinforce and correspond to our initial triple. We must take issues such as different concept spellings and concept synonyms (2), feature-synonyms (3) and paraphrasing of identical features (4) into account.

2b. Relations

Relations pose similar problems, and are all the more complicated because in general there is such a wide variety of ways of describing the same relation (notwithstanding the number of synonyms for any given feature). Even if we restrict our feature synonym set to just "¬slow" and "fast", some sentences containing information about aeroplanes could include:

1. Aeroplanes are fast.
2. Aeroplanes fly fast. / Aeroplanes go fast.
3. It was an uncharacteristically slow aeroplane.
4. Aeroplanes are slow on the ground, but fast in the air.
5. There is no such thing as a slow aeroplane.

For example, the statements in 3 and 4 seem to contradict that in 5.

Thus deriving features and relations is by no means trivial, especially since basic real-world knowledge (e.g. *aeroplanes are fast*) is comparatively unlikely to appear in corpora. Few writers take the trouble to state the obvious, since they (correctly) assume their reader is already equipped with knowledge of the obvious.

3. Gold Standard

We use the 512 concepts (ranging from aeroplane to zebra, both living and non-living concepts) of McRae et al (2007) as our target list, to which we will compare the output of our system. Features and relations were collected from approximately 725 participants in the following manner:

- Participants were offered, for each concept, 10 blank lines to list statements describing the concept.
- Each concept had exactly 30 participants listing features for it.
- Obviously synonymal features/relations were grouped together but conservatively to preserve any/all distinctions between meaning.
- Only features/relations which were cited by at least five individuals were included as targets.

4. Information Acquisition

We will employ a large corpus of text to "learn" our triples, using a variety of computational linguistic methods. For example the Gigaword corpus contains around 1.2 * 10^9 words of English text from New York Times, Associated Press and other newswire sources.

We propose two possibilities for acquiring the types of relations and features we seek:

4a. Word Space Models

The underlying idea of word space models is to represent the semantic content of words through their cooccurrence patterns within a large body of text, i.e. how often other words occur "close" to them in text.

In this way, each word is represented as a high-dimensional vector whose dimensions correspond to (a function of) its co-occurrence scores, with each dimension representing a specific word. Similarity and relationships between concepts can then be derived using standard geometrical vector similarity measures — the logic being that similar words will demonstrate similar cooccurrence patterns and thus similar vectors.

In this way we hope to derive the features from closely related words within the vector space. Deriving the relations will be more difficult (because of their sometimes multi-word nature) but we would hope to employ more advanced three-way vector similarity methods to deal with this.

4b. Pattern-based text matching

Another possible method is one which employs simple pattern searching, using an initially partially annotated corpus, where **concept** and **relation** are known, or where **concept** and **feature** are known. For example, supposing we knew that *an aeroplane is used for transportation*. We could then search for:

*An aeroplane is used for **

where we would want the * to have the same format as *transportation* (in this case, a noun). Similarly we could try to find instances of:

*An aeroplane * transportation*

where we would want the * to contain phrases filling a similar structure to *is used for* (e.g. past participle of a verb, followed by *for* as a preposition).

5. Discussion

We are still in the early stages of developing our system, but our preliminary results seem promising. By running statistical analyses on our output and evaluating with both the McRae norms and humans we hope to further enhance performance.

This project is unique in that it is drawing its model of how to represent feature/relations directly from theories developed in the field of cognitive neuroscience.

We are working closely with are colleagues there to improve the overlap between the two fields, and hopefully develop a model of conceptual structure which is applicable both to computer science and cognitive neuroscience.

References

Ken McRae, George Cree, Mark Seidenberg, Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. In *Behavior Research Methods, Instruments, & Computers* 2005, 37 (4), 547-559.

UNIVERSITY OF CAMBRIDGE