COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2013) 1–45
Copyright © 2013 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12091

Automatic Extraction of Property Norm-Like Data From Large Text Corpora

Colin Kelly,[a] Barry Devereux,[b] Anna Korhonen[a]

[a]*Computer Laboratory, University of Cambridge*
[b]*Department of Psychology, Centre for Speech, Language and the Brain, University of Cambridge*

Received 13 June 2011; received in revised form 1 April 2013; accepted 8 April 2013

Abstract

Traditional methods for deriving property-based representations of concepts from text have focused on either extracting only a subset of possible relation types, such as hyponymy/hypernymy (e.g., ***car*** *is-a* **vehicle**) or meronymy/metonymy (e.g., ***car*** *has* **wheels**), or unspecified relations (e.g., ***car***—**petrol**). We propose a system for the challenging task of automatic, large-scale acquisition of unconstrained, human-like property norms from large text corpora, and discuss the theoretical implications of such a system. We employ syntactic, semantic, and encyclopedic information to guide our extraction, yielding concept-relation-feature triples (e.g., ***car*** *be* **fast**, ***car*** *require* **petrol**, ***car*** *cause* **pollution**), which approximate property-based conceptual representations. Our novel method extracts candidate triples from parsed corpora (Wikipedia and the British National Corpus) using syntactically and grammatically motivated rules, then reweights triples with a linear combination of their frequency and four statistical metrics. We assess our system output in three ways: lexical comparison with norms derived from human-generated property norm data, direct evaluation by four human judges, and a semantic distance comparison with both WordNet similarity data and human-judged concept similarity ratings. Our system offers a viable and performant method of plausible triple extraction: Our lexical comparison shows comparable performance to the current state-of-the-art, while subsequent evaluations exhibit the human-like character of our generated properties.

Keywords: Natural language processing; Property norm; Wikipedia; Human evaluation; WordNet; Pointwise mutual information; Log-likelihood; Entropy

Correspondence should be sent to Colin Kelly, Computer Laboratory, University of Cambridge, 15JJ Thomson Ave., Cambridge CB3 OFD, UK. E-mail: colin.kelly@cl.cam.ac.uk

1. Introduction

Humans' mental representation of the world is founded, in part, on concrete concepts such as ***car***, ***zebra*****,** and ***banana***. The nature of how these representations manifest and express themselves in the mind has been studied extensively in cognitive science, and recent theories of conceptual representation have adopted a distributed, componential, and feature-based paradigm (e.g., Farah & McClelland, 1991; Randall, Moss, Rodd, Greer, & Tyler, 2004; Tyler, Moss, Durrant-Peatfield, & Levy, 2000). According to these accounts, concepts are exhibited as patterns of activation across interconnected feature nodes (e.g., *has* **wheels**, *has* **stripes**, *has* **skin**). An important perceived advantage of such models is that they are able to naturally reflect a number of the desirable qualities of a conceptual representation framework. For example, semantic similarity can be intuitively described by way of overlapping patterns of activation, which have been shown to offer predictions consistent with empirical evidence of semantic priming effects (Masson, 1995).

To test these theories, cognitive psychologists (e.g., Randall et al., 2004; Cree, McNorgan, & McRae, 2006; Grondin, Lupker, & McRae, 2009) have recently moved toward employing empirically grounded, real-world knowledge to instantiate their models of conceptual representation. To date, such knowledge has principally been derived from property norming studies in which a large number of participants write lists of properties (or "norms") of concepts. McRae, Cree, Seidenberg, and McNorgan (2005) collected such a set of norms, which we call the "McRae norms." Such norming studies have been used for implementing and testing models of conceptual representation, experimenting with various accounts of distributed conceptual knowledge in psycholinguistic studies (e.g., McRae, De Sa, & Seidenberg, 1997; Randall et al., 2004; Cree et al., 2006; Tyler et al., 2000; Grondin et al., 2009).

The McRae norms broadly fall into a ***concept*** *relation* **feature** triple pattern,[1] which usually takes the form <***noun***> <*verb*> <**noun/adjective**>. These norms contain a wide variety of information types such as location (***knife*** *found in* **kitchens**), color (***cherry*** *is* **red**), parts (***cup*** *has* **a handle**), and uses (***ladle*** *used for* **stirring**). A significant minority of the properties are "behavior" properties, expressing activities often or typically undertaken by the concepts. These behavior properties do not take a *relation* verb (instead their relation is labeled *beh*) and thus usually take the form <noun> *beh* <**verb**> (e.g., ***airplane*** *beh* **crashes**, ***airplane*** *beh* **flies**).

Data from property norming studies suffer a number of shortcomings (which have been examined extensively in the literature; e.g., Murphy, 2003; McRae et al., 2005). One such weakness is that participants often under-report certain properties, even when they are facts presumably known by the participants. For example, although *is* **animal** is listed as a property of the majority of animals appearing in the norms, *beh* **breathes** is listed only as a property for ***whale***. Similarly, *has* **heart** is not reported as a property for any animal concept even though all participants are likely to have known that animals have hearts. A related issue is inconsistency across highly related concepts: Although *has* **legs** is listed as a property of ***leopard***, it is absent for ***tiger***. Furthermore, participants are only able to report properties which they can put into words: Their mental representation of concepts is far richer than that which they are able to verbalize. This productive (verbalizable) set of properties constitutes only a relatively small subset of a participant's receptive property-set (i.e., those properties which the participant would deem to be correct but may not verbalize).

The objective of our work was to emulate and complement such norming studies by creating a system capable of automatically extracting these types of properties, using techniques from Natural Language Processing (NLP). The ability to do this would be of enormous benefit to experimental psychologists: It would enable them to avoid the labor-intensive task of manually generating new norms and would allow them to perform large-scale experiments using property norm data for any concept(s) of their choosing (no longer relying on a pre-determined set of normed concepts).[2]

It is important at this stage to draw out a theoretical distinction between the various types of semantic data that exist. Humans make use of three principal sources of semantically meaningful data: data already in the mind (conceptual knowledge), data in language (both spoken and written), and data in the world (i.e., data which we derive from our non-language-based experiences in, and interactions with, the world). The conceptual data already in the mind (if we ignore the possibility of innate semantic knowledge at birth) will have been derived by way of the other two sources, but it may also combine with them to produce further conceptual knowledge (inferences).

Property norms could be viewed as a window to this conceptual knowledge (albeit a window which does not encapsulate all of the conceptual knowledge, as demonstrated by the shortcomings listed above). However, for concrete nouns at least, property norms could also be viewed as a mere transcription of "data in the world": the norms listing real-world properties of real-world concepts. That the norms are, by necessity, written in language introduces the notion of them also being in the domain of "data in language"; property norms lie at the crossroads of the three data sources, which is perhaps why in previous work in cognitive psychology the various sources have sometimes been conflated.

There is clearly overlap between the three sources of data, and there are key theoretical questions of how similar they are (modulo the data's representation from each source) and the degree of overlap between them. Our work strives to go some way toward answering the question of whether the data in language (for which we use our corpora as a proxy) is a sufficient vehicle to capture the full scope of conceptual knowledge; can we, given a sufficiently large body of text, generate all the conceptual knowledge that could be found in the human mind? We will return to these questions in our discussion, but for now we review previous work in this area. We begin by questioning whether what we are aiming to do is in fact realistic, in terms of the extent to which conceptual knowledge may be solely extracted from text corpora. Louwerse (2010) argued for an "interdependency account," hypothesizing that meaning is embodied by the combination of both perceptual components and symbolic components (represented by computational/distributional approaches) which are mutually reinforcing. Following from this, Johns and Jones (2012) proposed a model which integrates perceptual information with the exposure to regularities in language as the foundations for the learning of lexical semantic representations. Indeed, Bruni, Tran, and Baroni (2011) instantiated such a model by combining classical text-based distributional data with image data as the perceptual information source and showed that the inclusion of the latter led to qualitative differences in performance.

These hypotheses of two distinct and separate types of data are notable, since in our work we hope to create a system able to extract perceptual/experiential data directly from distributional data.

Directly related to this article, Andrews, Vigliocco, and Vinson (2009) similarly proposed a theory of semantic representation based on a statistical combination of "experiential" data ("derived by way of our experience with the physical world") using property norms and "distributional" data (which "describes the statistical distribution of words across spoken and written language") from the British National Corpus (BNC). They were among the first to do so, stating that in previous literature the contributions of these two types of data had only been considered independently, never simultaneously or in combination. They evaluated their experiential and distributional models independently and in combination using six data sets offering semantic similarity measurements and found that their coupled model outperformed both of the constituent models alone. We believe this result motivates the use of pre-existing property norm data to guide our search for further properties. Even though their results could imply that not all the information we seek lies in text corpora, we note that they have not used syntactic knowledge in any of their models to date (although they have more recently employed word-order information in the framework of hidden Markov Models to further enhance their model; Andrews & Vigliocco, 2010). In a similar vein, Steyvers (2010) augmented probabilistic topic models derived from a text corpus with information from semantic property norms, generating "feature-topics" which were able to predict missing words in documents, again motivating the combination of a corpus and known property norms to discover semantic information.

The primary purpose of Natural Language Processing is to enable computers to handle text intelligently and to understand natural human language. The main aim of our work was to directly emulate property-norm-like relationships between concepts by using NLP techniques. However, we note that even from a theoretical perspective, one of the key benefits of using computational techniques to achieve this is that our output will not be limited by the number of humans generating a finite set of properties for each concept. We can rather, in theory, extract an extremely large number of properties for a given concept, presumably with varying degrees of relevance and specificity; it is this quantification of the relative relevance/salience of the extracted properties which we view as one of the key benefits of such a system. We believe it unlikely that humans' mental representation could be encapsulated in something as black-and-white as a short list of true properties (implicitly to the exclusion of any other properties), and any system which broaches this task should presumably take this into account; the ultimate output of such an idealized system will thus be inherently different from that produced by humans, and we will further discuss the theoretical implications of this later in this paper. However, for now we focus on our present work, the techniques of which we view as one possible route toward creating such a system.

Early work on computational semantics focused on finding semantically similar or related words: One of the first hypotheses in NLP was that word meaning could be modeled as a high-dimensional vector, where the vectors are derived from a corpus. For example, Deerwester, Dumais, Furnas, Landauer, and Harshman (1990) created their seminal Latent Semantic Analysis (LSA), which works by deriving such vectors from collections of discrete, segmented documents, creating a normalized co-occurrence matrix in which rows correspond to words and columns to documents and reducing the dimensionality of these matrices using Singular Value Decomposition. This enables the computation of the similarity of two words using the cosine angle between their reduced-dimensionality vectors. More recently, research has moved in the direction of discovering the precise and explicit nature of the relationships between words (and the concepts which they represent).

In the field, there is already a significant body of work on the topic of relation extraction. As property norming studies aim to gather relationships between entities from humans, relation extraction aims to automatically extract relationships between entities from text corpora. Hearst (1992) proposed a lexico-syntactic pattern-based approach for the automatic extraction of hyponyms, and many others have built on her ideas in a variety of ways. For example, relation extraction has been used for ontology learning: Rindflesch, Tanabe, Weinstein, and Hunter (2000) employed NLP methods to extract relationships between cancer therapy genes and drugs from a database of biomedical abstracts. Pantel and Pennacchiotti (2008) were able to extract specific semantic relations (e.g., *is-a*, *part-of*) from text and link them into existing semantic ontologies. A key difference between our work and that of ontology learning is that we seek common-sense properties of a prototypical nature, whereas ontologists are in search of scientific truths. For example, the McRae norms state that ***tomato*** *is* **vegetable**, while from a botanical perspective it is a fruit. Such patterns have also been used for tasks such as named-entity classification, where input strings are classified into "Person," "Organization," or "Location" classes. Collins and Singer (1999) used unsupervised earning of relations for this task. In named-entity extraction or ontology-learning the relationships and entities are usually well defined: The classes of the words/relations that are sought are closed and there is therefore less ambiguity about whether a particular entity/relation is valid. This relative consistency makes their appearances in corpora more predictable and renders detection and extraction a far easier task, with relatively good performance achievable through shallow methods (i.e., methods not requiring deeper syntactic/semantic information, such as part-of-speech or grammatical dependency data; Etzioni et al., 2005; Poon & Domingos, 2010). Finally, there has been some work in "generic relation" extraction: Davidov, Rappoport and Koppel (2007) and Davidov and Rappoport (2008) attempted to extract novel relationships rather than focusing on a fixed set of known relations. Their work does not, however, attempt to define the explicit nature of the relationships in question. Rather, it merely asserts their existence in the context of other, similar relations.

As already mentioned, the relations and features that we aim to extract are quite different in nature to those sought in ontology learning or named-entity recognition. For one, they are completely unconstrained, in that there are no limitations on what constitutes a semantic property of the kind that might appear in the McRae norms. This can yield rather long relations such as ***bayonet*** *found at the end of* **a gun**. Furthermore, relations can be idiosyncratic: One example is ***bomb*** *dropped from* **aeroplanes**—it is the only norm in the McRae set which has the word "dropped" as its relation. The types of relations are also highly concept dependent; for example, animal concepts will typically have visual properties (e.g., properties for ***zebra*** include *has* **stripes**, *has* **tail**, *has* **hooves**), while tools will usually have properties describing their make-up and functions (***umbrella*** has *used for* **keeping dry** and *made of* **plastic** listed as properties). Unconstrained property discovery is notoriously difficult, and consequently our task is not by any means an easy one.

The majority of prior work on finding properties of concepts has been limited in scope in order to make the task more workable: Barbu (2008) used shallow methods, employing a combination of pattern matching and association strength, to predict property norm-like descriptions for concepts. However, his method was unable to explicitly postulate new relations for the concepts, rather it predicted which semantic class (e.g., "Location," "Part," "Action") the extracted features belonged to. It did, nevertheless, offer the insight that not all features are created equal, and class-dependent methods may be key to extracting features (the pattern-based approach appeared to work well for certain semantic classes). Almuhareb and Poesio (2004, 2005) compared four vector-based concept description models, but only applied their work to clustering nouns into WordNet-derived classes. The main issue with their method was that the output did not yield a human-interpretable property-based description of a particular concept: The vectors they produced contained thousands of entries. Comparing these vectors may offer a good indication of the extent to which two concepts are similar (in terms of how well they cluster together), but inspecting them does not make explicitly clear the psychological properties which we seek. Nonetheless, we believe their results motivate the use of parsed text to improve concept descriptions.

In recent years, researchers have begun to develop methods which can automatically extract property norm-like representations from corpora. The development of sophisticated methods for the extraction of these relations necessitates a much deeper analysis of texts. It also requires a solid grasp of the linguistic phenomena that characterize the conceptually motivated properties we seek. Indeed, Baroni and Lenci (2010) use parsed text to instantiate their "Distributional Memory" framework, which is designed to act as a generalized distributional model of language suitable for a number of NLP tasks (including property extraction). However, only Baroni and Lenci (2008), Baroni, Murphy, Barbu, and Poesio (2010), and Devereux, Pilkington, Poibeau, and Korhonen (2009) have attempted to broach the specific and ambitious task of attempting to automatically generate property norm-like data.

Baroni et al. (2010) introduced an alternative approach to word space models called Strudel, which searched for "semantically meaningful patterns," rather than merely flat co-occurrence of words. It took a list of concepts and a part-of-speech tagged corpus, and searched for those nouns, verbs, and adjectives which were linked to the target concepts by a finite set of "connector patterns," or templates. How these templates were defined (e.g., target and property are adjacent; linked by a possessive; connected by a preposition) dictated what types of relationships were recorded. The second step of their system ranked the concept-property pairs based on the number of distinct linking patterns (rather than by the frequency of pattern instances). This involved grouping concepts and properties with similar patterns and creating shallow descriptions of them, noting the distribution of the generalized patterns (based on the sequences of part-of-speech patterns and their relative distribution in the corpus), which they called "type sketches." Of the models they tested, the Strudel method gave highest precision.

Devereux et al. (2009) presented a two-stage large-scale property extraction system, which employed class-based semantic information to guide its extraction, and was the first such method to focus not only on finding features of concepts but also on predicting the relation labels between those concepts and features. The first stage (the candidate feature extraction step) of this method was relatively simplistic: The RASP parser (Briscoe, 2006) was used to generate a list of grammatical relations (GRs; GRs describe linguistic relationships between words in a sentence, e.g., a direct object relation). Using this list, one could follow paths (beginning at the concept in question) indicated by the list to generate candidate features of that concept. Any and all nouns and adjectives path-linked to the concept by way of a non-auxiliary verb were considered potential features. The linking verb was returned as the corresponding "relation" for the candidate concept-relation-feature triple. Their relation/feature extraction did not take into account lexical or syntactic constructions that would typically be suggestive of features of the type we are aiming to extract. Their second stage (the feature reranking step) employed class-based semantic information to upweight semantically relevant features. However, it did not employ a variety of other measures which could similarly improve performance, such as those employed in the Baroni and Lenci (2010) method. Furthermore, the two stages were conducted independently from one another—syntactic information acquired from the parsing was lost in the subsequent reranking stage. This system acts as the baseline method for our experiments.

In our work, we use syntactic, semantic, and encyclopedic information to guide our extraction of unconstrained concept-relation-feature triples. We propose a novel system which generates candidate triples using grammatical relations extracted from two corpora: one encyclopedic (Wikipedia) and one general (the British National Corpus). We parsed both our corpora using the C&C parser, a fast and accurate parser (Clark & Curran, 2007a, b).

For each sentence, the C&C parse generates a list of (usually binary) grammatical dependencies between the constituent words. From this list, we may construct a GR graph representing the grammatical structure of that sentence. Using a series of rules, we select those paths through the graph containing relations and features which are likely to approximate property-based conceptual representations. Our rules take into account such information as the nature of the GRs in the path, the part-of-speech tags of the concept, relation and feature, as well as path-length information. We place an emphasis on ensuring the relations/features we extract are linguistically motivated. The system subsequently weights and ranks the extracted triples by way of a linear combination of four statistical metrics, selected to maximize the possibility that higher ranked properties would emulate human-generated norms. For example, we choose to downweight properties shared across a very large number of concepts, as these extremely common (and therefore highly general) features are unlikely to be cited as properties for any concepts (e.g., *be* **used**, *have* **thing**). As already mentioned, our ultimate aim is to go beyond human-elicited norms and extract a full picture of each concept through its properties. That said, we would still want to avoid (or at least downweight) very "general" properties, that is, those which are arguably common to all concepts (e.g., ***penguin*** *do* **exist**, ***car*** *be* **thing**). It is relevant and specific properties which we are interested in, at least in the first instance. We estimate the strength of association between the concepts and features extracted, hypothesizing that a higher degree of association will correlate well with human-like norms. We also make use of information directly acquired from the property extraction stage to guide the reweighting, forging a link between the candidate property extraction stage and the reweighting stage. Finally, we also employ the "semantic reweighting factor" as described by Devereux et al. (2009) as a parameter in our reweighting.

The challenging nature of our task is compounded by the difficulties encountered when attempting to evaluate the system output—the lists of features gained from property norming studies are usually non-exhaustive and often inconsistent across concepts, for example, *has* **eyes** is listed as a property of some animals but not others, and as such there is no true "gold standard" for evaluating our work. We therefore also need to employ additional evaluation techniques capable of addressing these problems. In total, we use three evaluation methods: standard NLP comparisons with a gold standard derived from the McRae norms, a comprehensive human-based evaluation, and a novel semantic-similarity vector space method which compares our output to that generated by 10 humans' judgments of semantic similarity.

In summary, this article addresses the challenge of extracting unconstrained human-like, common-sense property norms from large text corpora, as well as the theoretical implications of this task. Such norms have been used extensively in experimental psychology research (e.g., in work on concepts, categorization and semantic memory). The ability to automatically and accurately extract such properties could prove extremely beneficial for any researchers employing property-norm information in their investigations, by removing the limitations imposed by a fixed and restricted set of norms. Our proposed system employs syntactic, semantic, and encyclopedic information to guide the extraction of concept-relation-feature triples. We illustrate the value of this information during our candidate feature extraction stage and demonstrate the ability of statistical metrics to upweight more human-like features.

The primary contributions of our work are multiple. We employ only a relatively small set of rules to extract candidate concept-relation-feature triples and introduce a new measure, relation entropy, for gauging the strength of a candidate triple based on the number of rules yielding that triple. We reweight our system's candidate features in a novel way, using Entropy, Pointwise Mutual Information, and Log-likelihood Ratio measures combined with semantic information and compare how encyclopedic and general corpora (Wikipedia and the BNC) produce different "types" of triples. Our system also aims to extract behavioral features specifically exhibited by the concept under consideration, rather than activities merely associated with the concept at hand (an issue which previous systems have not broached). Our evaluation analyses demonstrate consistently comparable performance with respect to previous work in the same domain and we offer a comprehensive evaluation of our system: We compare it directly with a gold standard, we ask humans to evaluate its output manually, and we measure our system's capacity for predicting both WordNet and human-rated similarities between concepts. Finally, we discuss the theoretical implications of such property extraction in detail, situating the potential output of our computational linguistic techniques within the broader domain of conceptual and semantic knowledge in cognitive science.

2. Method

2.1. Property norms

McRae et al. (2005) collected a set of norms listing properties for 541 concrete concepts. In that study, the properties listed by different participants were normalized by mapping property descriptions with identical meanings to the same property label.[3] In this way, a list of normalized property labels was constructed, with a production frequency value (i.e., the number of participants who produced a specific property for a given concept) associated with each concept-property pair. Other property norming studies exist (e.g., Garrard, Ralph, Hodges, & Patterson, 2001; Vinson & Vigliocco, 2008; Devlin, Gonnerman, Andersen, & Seidenberg, 1998), but we focus our experiments on the McRae set because it is the largest to date, offering a comprehensive set of norms to train and evaluate our system on.

For the purposes of our experiments, it was necessary to recode the relatively free-form McRae norms into a more coherent and structured representation to allow for easier and more rigorous computational manipulation, as well as to facilitate evaluation. Each concept's properties were converted into a ***concept*** *relation* **feature** format (e.g., ***bat*** *have* **wing**) as follows: If the property is a behavior, then the infinitive of the verb acts as the feature and the verb *do* acts as the relation. Otherwise, the final noun/adjective of the property is employed as the feature and the main (non-auxiliary) verb used as the relation. We remove all determiners: The vast majority of the norms contain general properties, with very few containing determiners whose specificity is likely to significantly alter a given property's meaning. We also remove prepositions: Although we accept that there is a sizeable minority of the norms—typically functional properties—for which prepositions can carry meaning (e.g., the relations *used-for*, *used-by*, *used-with,* and *used-in* are clearly semantically different), we decided that, at this stage, evaluating our system without prepositions would offer a better overall picture of its performance. Next, if the property contained an adjective-noun combination, this would be recoded and split into two separate concept-relation-feature triples. Although this separation of concept properties into constituent key sections (and omission of certain aspects of some of the original properties) is a simplification, the amount of information lost is actually relatively small—in the vast majority of cases, the three-term triples returned are true to their original meaning. Table 1 gives the ten most frequent (normalized) features for two concepts in the norms, ***car*** and ***penguin***, and their corresponding recoded triples.

2.2. Extraction method

Our system is outlined in Fig. 1. The input to our system consists of (a) the set of concepts for which we aim to find properties and (b) C&C-parsed sentences from our chosen corpus. The system works in two main stages, which we summarize briefly now, and which will be explained in more detail in the sections that follow. In the *Extraction Rules* section, we will also give a worked example of how a triple such as ***penguin*** *have* **feather** could be extracted using our system.

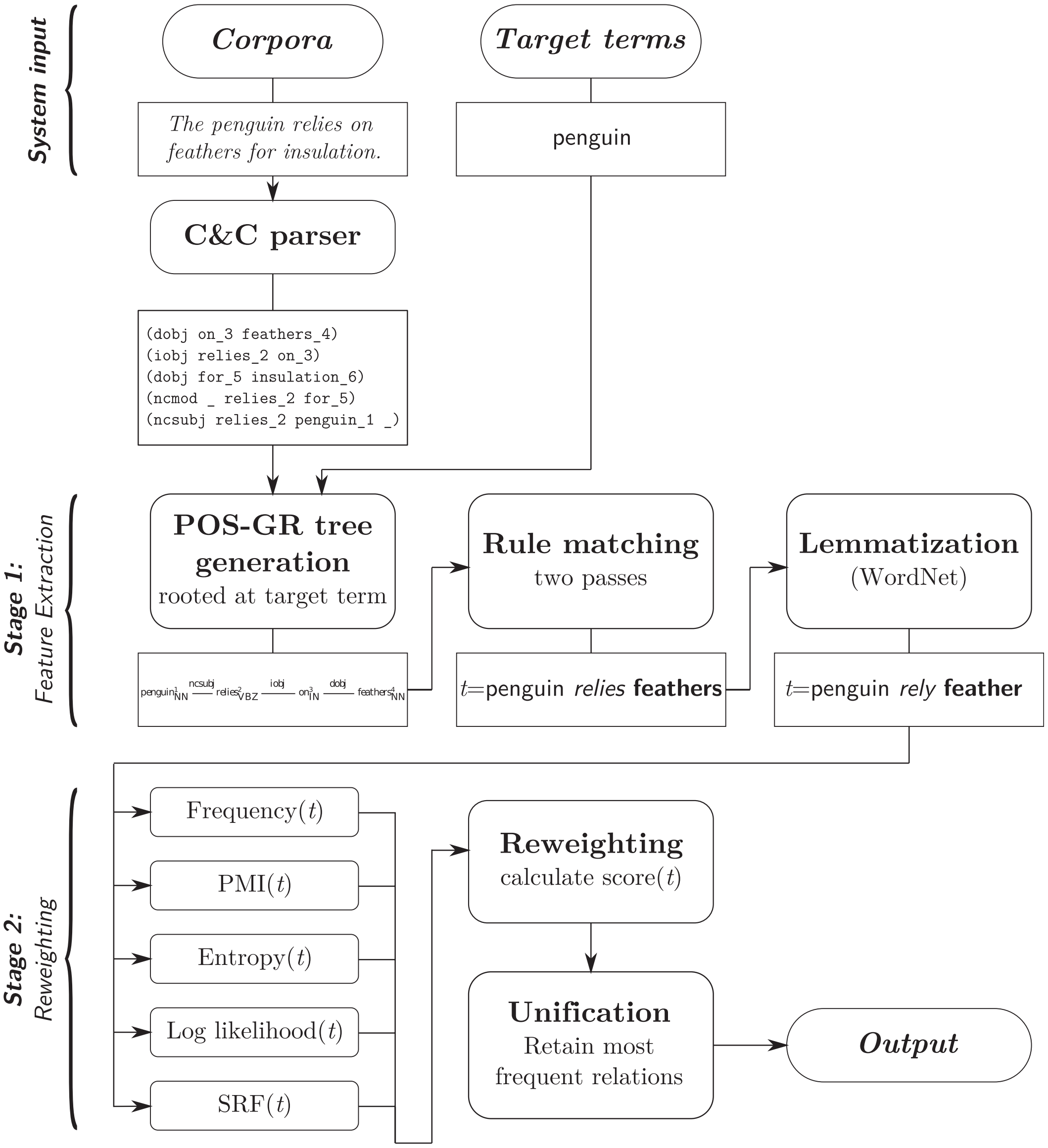
Our system is trained on 489 of the McRae concepts and will subsequently be evaluated on the 44 remaining concepts.

2.2.1. Corpora

We employ three corpora for our experiments: two are subsets of Wikipedia (the Wiki500 and Wiki100K corpora), and the other is the British National Corpus (Leech,

Table 1
Sample features from the McRae norms with their frequency and our corresponding recoded concept-relation-feature triples

	car	
McRae feature	Recoded triple	Freq
has wheels	***car*** *have* **wheel**	19
used for transportation	***car*** *use* **transportation**	19
has 4 wheels	***car*** *have* **4 wheel**	18
has doors	***car*** *have* **door**	13
has an engine	***car*** *have* **engine**	13
requires petrol	***car*** *require* **petrol**	12
has a steering wheel	***car*** *have* **steering wheel**	12
used for passengers	***car*** *use* **passenger**	9
a vehicle	***car*** *be* **vehicle**	9
is fast	***car*** *be* **fast**	9
	penguin	
is black	***penguin*** *be* **black**	24
a bird	***penguin*** *be* **bird**	22
is black-and-white	***penguin*** *be* **black-and-white**	22
is white	***penguin*** *be* **white**	22
has a beak	***penguin*** *have* **beak**	21
beh—cannot fly	***penguin*** *cannot* **fly**	20
beh—waddles	***penguin*** *do* **waddle**	15
beh—swims	***penguin*** *do* **swim**	14
lives in cold climates	***penguin*** *live* **climate**	13
has feet	***penguin*** *have* **foot**	12


System input
Corpora
Target terms
The penguin relies on feathers for insulation.
penguin
C&C parser
Stage 1: Feature Extraction
POS-GR tree generation rooted at target term
Rule matching two passes
Lemmatization (WordNet)
$\text{penguin}^{1}_{NN} \xrightarrow{\text{ncsubj}} \text{relies}^{2}_{VBZ} \xrightarrow{\text{iobj}} \text{on}^{3}_{IN} \xrightarrow{\text{dobj}} \text{feathers}^{4}_{NN}$
t=penguin *relies* **feathers**
t=penguin *rely* **feather**
Stage 2: Reweighting
Frequency(t)
PMI(t)
Entropy(t)
Log likelihood(t)
SRF(t)
Reweighting calculate score(t)
Unification Retain most frequent relations
Output

```
(dobj on_3 feathers_4)
(iobj relies_2 on_3)
(dobj for_5 insulation_6)
(ncmod _ relies_2 for_5)
(ncsubj relies_2 penguin_1 _)
```

Fig. 1. An overview of our system, outlining the system input and two main stages: "Feature Extraction" and "Reweighting."

Garside & Bryant, 1994). The Wiki500 corpus (1.1m words) contains about 500 Wikipedia articles corresponding to each of the McRae concepts, while the Wiki100K corpus (36.5m words) comprises those Wikipedia articles with titles containing one of the McRae concepts (and with a title-length of five words or less). For a full description of how the Wikipedia subcorpora were generated, see Devereux et al. (2009). We chose Wikipedia because it forms a large, comprehensive, and freely accessible source of encyclopedic knowledge, and we are confident that much of the property norm information we seek is likely to be encoded within it. Indeed, nearly all the McRae concepts have their own articles within Wikipedia, and generally the majority of cited properties for a given concept can be found in its Wikipedia article, albeit rarely expressed in an identical (or indeed similar) way to the property norms.

We also employ the 100-million-word British National Corpus (BNC) which contains a sample of written (90%) and spoken (10%) UK English collected from 1960 to 1993 (Burnard, 2007). It is balanced across domains in that it is not limited to any particular subject, genre, or field and is designed to represent a broad cross section of modern British English.

These corpora were chosen to illustrate the extent to which the types of property norm-like information we seek can be found in both encyclopedic and general contexts. Although we might expect much of the human-produced knowledge found in the McRae norms to also exist in our encyclopedic resource—implying a certain degree of completeness to Wikipedia with regard to our task—this is not always the case. For example, the triple *eaten by* **monkeys** appears as a property of ***banana*** in the McRae norms, but the word ‘monkey’ does not appear at all in the Wikipedia *banana* article. Hence, we also hope to assess the extent to which those properties not included in Wikipedia (perhaps due to their common-sense rather than essential nature) might instead be encoded in everyday speech and text, such as that contained in the BNC. We will later compare the properties returned from each corpus to investigate whether they complement one another, as this would motivate using a combination of the two.

2.2.2. Parser

For our experiments, we employed the C&C parser (Clark & Curran, 2007a, b) to extract both grammatical relations (GRs) and part-of-speech (POS) tags from the sentences within our corpora. GRs denote the functional relationships between different words within a sentence, offering a structured representation of the underlying grammatical organization of a given sentence. The RASP parser (Briscoe, 2006) has been used in previous work (Devereux et al., 2009; Kelly, Devereux, & Korhonen, 2010), but we have chosen C&C because it has been shown to have better parser accuracy over RASP overall (Clark & Curran, 2007a). Specifically, it has also been shown to outperform RASP on a majority of the grammatical relation types that we will employ in our rules (e.g., direct and indirect objects, non-clausal modifiers and subjects). It is also a lexicalized-grammar parser: It takes into account surface-level lexical information when predicting part-of-speech tags and grammatical dependencies, something RASP ignores, and it parses text much more quickly than RASP. We parsed all three of our corpora using C&C.

2.2.3. Corpus processing

For each sentence in the corpus, the C&C parse output contains a set of GRs which form an undirected acyclic graph whose nodes are labeled with words from the sentence. It also contains POS information for each word in the sentence. It is possible to construct a GR-POS graph rooted at our target term, with POS-labeled words as nodes, and edges labeled with GRs linking the nodes to one another. Thus, the GR-POS graph interrelates all lexical, POS, and GR information for the entire sentence.

Our system executes two passes over the corpus. The first pass is designed to extract a list of strongly associated words as potential features for each concept to be used as input into one of our more noisy rules (Rule 8) in the second stage. This list of associated features is obtained by extracting those terms linked to the concept by short grammatical relation paths through the GR-POS graph (i.e., finding modifying nouns and adjectives, indirect objects and possessives relating to the concept). For example, sentences including phrases such as *it attacked the penguins' eggs* and *a penguin egg was found* would indicate, through the possessive and noun-noun compound constructions, that **egg** is a potential feature of ***penguin***.

The second pass employs a manually compiled rule-set (which we describe in the next section) to conduct a breadth-first search over all directed paths rooted at the target concept, logging each time a rule is fired. This process generates candidate concept-relation-feature triples, as well as their frequency of instantiation (according to our rules) across the corpus.

One of the rules (Rule 8) takes a modifying noun for a concept and employs the verb linking it to the concept as the relation verb. The rule is relatively noisy: Although it often matches valid concept-relation-feature triples (e.g., ***banana*** *include* **fiber**, ***banana*** *be* **crop**), it also frequently extracts incorrect triples (e.g., ***banana*** *favor* **withdrawal**, ***banana*** *prohibit* **law**, ***banana*** *feature* **underground**). Therefore, for this rule we only allow candidate triples which include features already associated with the concept during the first pass.

2.2.4. Extraction rules

Our rules from the second pass were constructed by taking a sample of concepts and their corresponding features from the McRae norms and then examining sentences from the Wiki500 corpus containing a concept and one of its features. Those sentences containing an instantiation of a likely triple would have the path linking the concept and feature through their GR-POS graph examined for a pattern of GRs and POS tags which would be strongly suggestive of a true relation/feature. Provided this pattern was not subsumed by any pre-existing rules, a new rule would be generated from it. We note that our rules do not explicitly take negations along the path into account. We include an outline of all of our rules in Table 2. For an in-depth explanation of the various POS and GR tags referenced in the rules, see Jurafsky and Martin (2000) (Appendix C) and Briscoe (2006), respectively.

In addition to extracting ***concept*** *relation* **feature** triples, we also place an emphasis on extracting behavior properties, which appear throughout the McRae norms (e.g., ***penguin*** *do* **waddle**). This is similar to the model of Baroni et al. (2010), although our rules aim to extract behaviors explicitly exhibited by the concept at hand rather than actions merely associated with the concept. In other words, we are aiming to de-emphasize behaviors such as ***motorcycle*** *beh* **ride** and ***motorcycle*** *beh* **park** to focus on ***motorcycle***

Table 2
Our 12 rules with the maximum path length (M) the rule will apply to, frequency information of rule-firing on the Wiki500 corpus, a description of the rule itself, an example sentence which fires the rule and the resulting triple

ID	M	Freq	Rule	Sentence	Triple
1	1	1,060	Feature has a `VBG` (“being”) tag and is linked to concept by an `ncmod` (non-clausal modifier), `xmod` (predicative relation modifier), `cmod` (clausal modifier), or `pmod` (prepositional modifier) GR then relation is *do*	American bison grazing in Custer State Park in South Dakota.	***bison*** *do* **graze**
2	1	7,848	Feature is a verb and is linked to concept by a `ncsubj` (non-clausal subject relation) GR then relation is *do* (unless tag is `VBG` (“being”), in which case relation is *be*)	A chain is usually made of metal.	***chain*** *be* **metal**
3	1	16,232	Feature has a `NN` (common noun) or `JJ` (general adjective) tag and is linked to concept by a `ncmod` (non-clausal modifier) GR then relation is *be*	In mechanical clocks this is either a pendulum or a balance wheel.	***clock*** *be* **mechanical**
4	1	6,953	Feature has a `NN` (common noun) tag and is linked to concept by a `ncmod` (non-clausal modifier) GR then relation is *have*	Coconut water can be used as an intravenous fluid.	***coconut*** *have* **water**
5	3	4,445	Feature has a `JJ` (general adjective) tag and is linked to the rest of its graph by an `xcomp` (unsaturated VP complement relation) or `ncmod` (non-clausal modifier) relation then relation is *be*	Chains can also be decorative as jewellery.	***chain*** *be* **decorative**
6	3	3,650	Feature has a `VBN` (past participle verb) tag then relation is *be*	Carrot flowers are pollinated primarily by bees.	***flowers*** *be* **pollinated**
7	3	242	Feature has a `NN` (common noun) tag then relation is *be*	The cathedral often being a large building serves as a meeting place.	***cathedral*** *be* **building**
8	4	7,572	Feature has a `NN` (common noun) tag and is linked to its graph by a `ncmod` (non-clausal modifier) relation then relation is only verb in path to feature	A chain may consist of two or more links.	***chain*** *consist* **links**
9	4	2,166	Feature has a `NN` (common noun) tag and is linked to graph by an `xcomp` (unsaturated VP complement relation) tag then relation is verb in path closest to feature	Airport trains are trains within air ports that transport people between terminals.	***trains*** *transport* **people**

(continued)

Table 2. (continued)

ID	M	Freq	Rule	Sentence	Triple
10	4	6,181	Feature has a `NN` (common noun) tag and there is a linking node node with a `VVN` (past participle) tag and the feature is linked to that node by an `xcomp` (unsaturated VP complement relation) GR then relation is that node	The tiny pharaoh ant is a major pest in hospitals and office blocks.	***ant*** *be* **pest**
11	∞	11,449	Feature has a `NN` (common noun) tag and final GR is `dobj` (direct object) and penultimate GR in path is `iobj` (indirect object) then relation is the penultimate node	Alligators are native to only two countries: the United States and China.	***alligators*** *native* **usa**
12	4	1,128	Feature has a `JJ` (general adjective) tag and relation node is a verb then relation is that verb	Tigers for the most part are solitary animals	***tigers*** *be* **solitary**

beh **travel**, ***motorcycle*** *beh* **cruise,** and ***motorcycle*** *beh* **speed**; we would prefer the former relationships to be yielded as ***motorcycle*** *be* **ridden** and ***motorcycle*** *be* **parked**. This involved specifically creating rules which focused on the verbs in the sentence. Creating these rules was challenging: There are no noun or adjective features to anchor the verb, and it is therefore more difficult to rigorously ascertain which verbs are feature verbs, and which are just incidental given the context. Corpus-based distributional models do not usually incorporate such distinctions.

Following from this, when constructing our rules it was also important to take account of directionality in the GR-POS graph. Doing this allows us to distinguish between such phrases as "dog bites man" and "man bites dog"—an essential step for understanding the meaning of a given sentence, in addition to better dealing with passive verb formations. We hope that doing this will also serve to reduce the amount of noise produced in our system's output: Previous systems (Devereux et al., 2009; Kelly et al., 2010) have not taken this directionality into account.

In constructing the rules in this way, our overriding aim was to ensure that the features and relations extracted were of a high quality, and likely to be true.

By way of example, the following sentence is found in one of our corpora:

"The penguin relies on feathers for insulation."

The C&C parse for this sentence yields, along with POS tags for each word in the sentence, the following GR output:

```
(det penguin_1 The_0)
(dobj on_3 feathers_4)
```

```
(iobj relies_2 on_3)
(dobj for_5 insulation_6)
(ncmod _ relies_2 for_5)
(ncsubj relies_2 penguin_1 _)
```

From these, we may construct the grammatical relation graph with POS tags as shown in Fig. 2. One of our concepts is ***penguin***, so we may examine all paths through the tree rooted at the concept. In our example, we find that the path found in Fig. 3 activates one of our rules (Rule 11), yielding the interim triple ***penguin*** *relies* **feathers**.

2.2.5. Lemmatization

We employed the NLTK WordNet lemmatizer (Bird, 2006) to lemmatize all extracted features and relations, using part-of-speech information to group together various inflected forms. This allows us to manipulate semantically identical (or near-identical) words as a single term. The feature head is lemmatized as an adjective or a noun unless it is a behavior feature, while the relation head and behavior features are lemmatized as verbs.

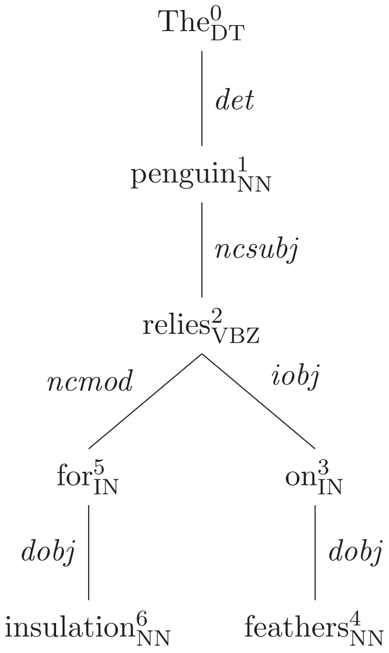

$\text{The}^{0}_{\text{DT}}$
det
$\text{penguin}^{1}_{\text{NN}}$
$ncsubj$
$\text{relies}^{2}_{\text{VBZ}}$
$ncmod$
$iobj$
$\text{for}^{5}_{\text{IN}}$
$\text{on}^{3}_{\text{IN}}$
$dobj$
$dobj$
$\text{insulation}^{6}_{\text{NN}}$
$\text{feathers}^{4}_{\text{NN}}$

Fig. 2. A C & C-derived grammatical relation tree for the sentence *The penguin relies on feathers for insulation.*

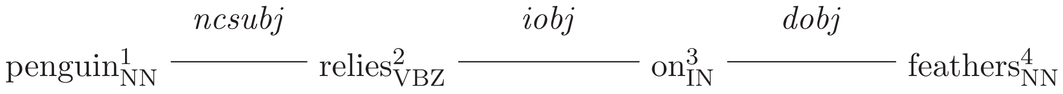

$ncsubj$
$iobj$
$dobj$
$\text{penguin}^{1}_{\text{NN}}$
$\text{relies}^{2}_{\text{VBZ}}$
$\text{on}^{3}_{\text{IN}}$
$\text{feathers}^{4}_{\text{NN}}$

Fig. 3. A path through our grammatical relation tree, activating Rule 11 to derive the triple ***penguin*** *rely* **feathers.**

For example, if we were to extract the triples ***bull*** *were* **cows**, ***balloon***—**floats** and ***cake*** *are* **best** they would be lemmatized to the form ***bull*** *be* **cow**, ***balloon*** *do* **float,** and ***cake*** *be* **good,** respectively. This stage is important for our evaluation, as it avoids the possibility of triples being marked as incorrect due to inflectional differences. Our example triple derived in the previous section (***penguin*** *relies* **feathers**) would thus be converted to ***penguin*** *rely* **feather**.

2.3. Reweighting metrics

Despite our efforts to ensure that the candidate triples we extract are plausible from a syntactic perspective, it is inevitable that some of them will be incorrect—this is because even grammatically identical constructions often have distinct meanings in different semantic contexts. Consider, for example, the sentences:

"Every cloud has a silver lining."

and

"Every beehive has a queen bee."

Both exhibit identical grammatical structure as demonstrated by their respective C&C parses. However, the first is an idiomatic phrase—no cloud possesses a literal silver lining—whereas the second shows a true property of beehives. Context and semantics often greatly affect the meaning of sentences within our corpora. We therefore expect the output of the first stage of our system to be quite noisy (i.e., producing incorrect triples). We need to ensure that the triples we select from our set of candidates are indeed likely to indicate correct semantic features and relations.

Therefore, the second stage of the system involves reweighting and selecting our output of relations/features from the first stage in a way that brings to the fore those which we might expect to find in property norm data. We employ four measures to achieve this and empirically test which linear combination of these four metrics yields the best results. Taking the linear combination in this way also allows us to assess the relative contribution of the metrics toward improving accuracy, illustrating the degree to which each of them is useful.

2.3.1. Pointwise mutual information

Pointwise mutual information (PMI) was first proposed by Church and Hanks (1990) as an objective measure for estimating word association norms. In information theory and statistics it is employed as a metric for measuring the strength of association between two events. It has been widely used in NLP as a measure of word similarity/semantic relatedness (Pantel & Lin, 2002; Turney, 2001). For our purposes, we will employ it as a measure of the strength of association between an extracted concept and its feature. In boosting those concept-feature pairs with high mutual information, we hope that more relevant and informative concept-relation-feature triples will come to the fore. For a given triple $t = (c, r, f)$, we calculate PMI as follows:

$$PMI(t) = \log \frac{freq(c,f) \times N}{freq(c) \times freq(f)} \text{ where } N = \sum_{i \in C} \sum_{j \in F} freq(i,j) \tag{1}$$

2.3.2. Entropy

We also calculate an entropy statistic for each extracted relation/feature pair based on its firing of rules during the second pass of our relation/feature extraction stage. If we define R_t as the set of rules which fire to produce a specific triple t, then we may define the entropy of t as follows:

$$\text{Entropy}(t) = -\sum_{r \in R_t} p(r|t) \log p(r|t) \tag{2}$$

where $p(r|t)$ is a probability mass function for the triple t across our rules. We calculate $p(r|t)$ as follows:

$$p(r|t) = \frac{p(r,t)}{p(r)} = \frac{freq(r,t)}{freq(t)}, \tag{3}$$

where $freq(r, t)$ is the number of times rule r fires to produce triple t and $freq(t)$ is the total frequency of triple t. In this way, $p(r|t)$ exhibits the usual properties of a probability mass function over the set of rules R_t.

We illustrate this with an example. Suppose we extract four triples (A–D) using four rules (r_1–r_4):

- *A*: ***dog*** *has* **tail** (generated from r_1 only, frequency 2)
- *B*: ***dog*** *is* **animal** (generated from r_1, r_2, and r_3, with frequencies 3, 2, and 5, respectively)
- *C*: ***dog*** *has* **bone** (generated from r_1 and r_3 with frequencies 3 and 2 respectively)
- *D*: ***dog*** *chases* **cat** (generated from r_4 only, frequency 3)

We therefore know the following:

$$\text{freq}(r_1,A) = 2 \quad \text{freq}(r_2,A) = 0 \quad \text{freq}(r_3,A) = 0 \quad \text{freq}(r_4,A) = 0$$

$$\text{freq}(r_1,B) = 3 \quad \text{freq}(r_2,B) = 2 \quad \text{freq}(r_3,B) = 5 \quad \text{freq}(r_4,B) = 0$$

$$\text{freq}(r_1,C) = 3 \quad \text{freq}(r_2,C) = 0 \quad \text{freq}(r_3,C) = 2 \quad \text{freq}(r_4,C) = 0$$

$$\text{freq}(r_1,D) = 0 \quad \text{freq}(r_2,D) = 0 \quad \text{freq}(r_3,D) = 0 \quad \text{freq}(r_4,D) = 3$$

And we also know that:

$$\text{freq}(A) = 2 + 0 + 0 + 0 = 2 \quad \text{freq}(B) = 3 + 2 + 5 + 0 = 10$$

$$\text{freq}(C) = 3+0+2+0 = 5 \quad \text{freq}(D) = 0+0+0+3 = 3$$

From these frequency values we calculate

$$p(r|t)$$

using Equation 3:

$$p(r_1|A) = 2/2 = 1 \quad p(r_2|A) = 0 \quad p(r_3|A) = 0 \quad p(r_4|A) = 0$$

$$p(r_1|B) = 3/10 = 0.3 \quad p(r_2|B) = 2/10 = 0.2 \quad p(r_3|B) = 5/10 = 0.5 \quad p(r_4|B) = 0$$

$$p(r_1|C) = 3/5 = 0.6 \quad p(r_2|C) = 0 \quad p(r_3|C) = 2/5 = 0.4 \quad p(r_4|C) = 0$$

$$p(r_1|D) = 0 \quad p(r_2|D) = 0 \quad p(r_3|D) = 0 \quad p(r_4|D) = 3/3 = 1$$

And we can next calculate the entropy for our triples:

$$\begin{aligned}
\text{Entropy}(A) &= -\sum_{r \in R_A} p(r|A) \log p(r|A) \\
&= p(r_1|A) \log p(r_1|A) \\
&= 1 \times \log(1)) \\
&= 0 \\
\text{Entropy}(B) &= -\sum_{r \in R_B} p(r) \log p(r|B) \\
&= p(r_1|B) \log p(r_1|B) + p(r_2|B) \log p(r_2|B) + p(r_3|B) \log p(r_3|B) \\
&= -(0.3 \times \log(0.3) + 0.2 \times \log(0.2) + 0.5 \times \log(0.5)) \\
&\approx 0.4471 \\
Entropy(C) &= -\sum_{r \in R_C} p(r) \log p(r|C) \\
&= p(r_1|C) \log p(r_1|C) + p(r_3|C) \log p(r_3|C) \\
&= -(0.6 \times \log(0.6) + 0.4 \times \log(0.4))
\end{aligned}$$

$$
\begin{aligned}
&\approx 0.2923 \\
\text{Entropy}(D) &= -\sum_{r \in R_D} p(r|D) \log p(r|D) \\
&= p(r_4|D) \log p(r_4|D) \\
&= -(1 \times \log(1)) \\
&= 0
\end{aligned}
$$

The logic behind taking account of and summing the scores over the number of rules fired is that if a relation/feature pair corresponds to multiple rules, then it is less likely to be a "false positive." As Baroni et al. (2010) observed, the phrases *the tail of the tiger* and *the tiger's tail* are both reasonably likely to occur relatively frequently in a sufficiently large corpus. If we contrast these with *the year of the tiger* and *the tiger's year*, we note that the former is quite likely to appear (due to the idiosyncratic and idiomatic meaning of *year of the tiger*), while we might expect *the tiger's year* to occur much less frequently than our other three phrases.

We wish to adjust the number of rules for a given feature by the probability that those rules will fire (rules which fire less frequently are presumably more difficult to satisfy, and, given that they were constructed specifically to extract accurate relations, they are presumably stronger predictors of an accurate triple). However, we also wish to prevent our less common rules from having a disproportionate effect on their upweighting ability due to their relatively low frequency of activation. Using the entropy curve allows us to take into account the incidence of rules firing relative to one another, while avoiding outlier rules (for a given concept) unduly influencing the scoring. This is a novel way of implementing Baroni et al.'s insight that triples derived from multiple rules or patterns are more likely to correspond to true properties.

2.3.3. Log-likelihood ratio

First proposed by Dunning (1993) for use in NLP, the log-likelihood ratio is a measure of the distribution of linguistic phenomena in texts. It has been used to contrast the relative frequencies of words in a corpus, and bring into relief lexical phenomena which are particularly distinctive in large bodies of text—for example, words which are under- or over-used compared to the norm. We employ the log-likelihood ratio across our set of concept-feature pairs (and their raw frequency data). Our aim was to derive those features which are particularly distinctive for a given concept, and which will therefore likely be features of that concept alone. Unlike PMI, the log-likelihood ratio has also been shown to work well under sparse data conditions (Dunning, 1993), making it particularly appropriate for our task.

We generate a frequency contingency table relating to each distinct concept-feature pair across our triples by grouping and summing the production frequencies of triples containing differing relations but the same concept and feature. For each concept-feature pair, this contingency table contains our observations across all triples of the (non-)occurrence of both the concept and the feature. For any given concept c and feature f, we define k_{11} to be the total frequency of concept c and f co-occurring across all triples, k_{12} to be the total frequency of triples with concept c but not with feature f. We define k_{21} as the total frequency of triples with f as feature, but without c as concept, and k_{22} as the total frequency of triples with neither c as their concept nor f as their feature. We then define the log-likelihood ratio for a given triple $t = (c, r, f)$ as follows:

$$LL(t) = 2\sum_{i,j} k_{ij} \log \frac{n_{ij}}{m_{ij}} \text{ where } n_{ij} = \frac{k_{ij}}{k_{i1} + k_{i2}} \text{ and } m_{ij} = \frac{k_{1j} + k_{2j}}{k_{11} + k_{21} + k_{12} + k_{22}} \quad (4)$$

2.3.4. Semantic reweighting factor

We employ the same method as described by Devereux et al. (2009) and Kelly et al. (2010) to assess the conditional probability of a feature given that it relates to a specific concept. We assume there is a two-dimensional probability distribution, $P(C, F)$, over concept classes, C (e.g., apparel, instruments), and feature classes, F (e.g., materials, activities). Knowing this distribution allows us to assess how likely it is that a candidate feature f is true for a concept c, assuming that we know that $c \in C$ and $f \in F$.

We estimated this probability distribution using the McRae norms, excluding our held-out ESSLLI set of concepts and their properties. We independently cluster both the concepts and feature-head terms from our non-ESSLLI recoded norms into two sets of 50 and 150 clusters,[4] respectively, using hierarchical clustering based on their similarity. Similarity between two words is defined as the maximum value of Lin's similarity metric (Lin, 1998) across all binary combinations of the two words' WordNet[5] senses. If f does not appear in the non-ESSLLI set of features, then f is simply assigned to the cluster it was most similar to. For example, one of our concept clusters contains (among others) the concepts ***asparagus***, ***blueberry***, and ***cranberry***; another contains ***ant***, ***bat***, and ***bear***. Similarly, one of our feature clusters contains **apple**, **berry,** and **carrot,** whereas another contains **accomplishment**, **achievement,** and **award**.

We use the production frequency information from the non-ESSLLI norms to estimate the probability distribution, $P(C, F)$, over all concept clusters, C, and feature clusters, F:

$$P(F|C) = P(C,F)/P(C) = \sum_{c \in C, f \in F} freq(c,f) / \sum_{c \in C} freq(c) \quad (5)$$

Then, for a given concept c and feature f we merely take the conditional probability of f, given the concept c, $p(f|c)$, to be $P(F|C)$, where C and F are such that $c \in C$ and $f \in F$. We define this conditional probability to be, for a given triple, t, $SRF(t)$, our Semantic Reweighting Factor.

2.3.5. Reweighting

To finally order the output of our system, we score and rerank the output triples (t = (c, r, f) where c, r, and f are a given concept, relation, and feature, respectively) as follows:

$$score(t) = \beta PMI \cdot PMI(t) + \beta Ent \cdot Entropy(t) + \beta LL \cdot LL(t) + \beta SRF \cdot SRF(t) + normfreq(t) \tag{6}$$

The multiplicands of the four free variables were normalized by finding, for each concept, the highest (`max`) and lowest (`min`) possible values of each metric across all triples $t \in T$, subtracting max from every triple and then dividing the result by the difference (max – min). In this way, the values of each metric lay between 0 and 1, with, for each concept, a maximum value of 1 for at least one triple and a minimum value of 0 for at least one triple. In cases where the difference was 0, all triples for that metric and concept were assigned zero (and consequently that metric, for that concept, would have no impact on the triple ranking).

Doing this allows a crude assessment of the importance of the metrics relative to one another. We also use a normalized frequency measure for each triple, normfreq(t). In this way, the trivial case when $\beta PMI = \beta Ent = \beta LL = \beta SRF = 0$ yields a ranking equivalent to ordering by frequency of extracted relations/features alone. We will offer empirically derived values for these parameters in the sections which follow.

2.3.6. Relation unification

The final part of the reweighting stage attempts to ascertain the most likely relation for similar triples. This involves concatenating all triples which share a concept and feature by removing the less frequent of these and summing their scores to the most frequent triple. For example, if we were to have extracted the triples ***penguin*** *splash* **water**, ***penguin*** *swim* **water**, ***penguin*** *live* **water** with scores 1, 4, and 5, respectively, then these would all be funneled into a single triple, ***penguin*** *live* **water**, with score 10. In our example above, ***penguin*** *rely* **feather** would be grouped under a triple such as ***penguin*** *have* **feather** (as this would likely be more frequent over the corpus).

3. Results

Having trained our system, we employ a number of evaluation methods to ascertain the quality of our extracted features and relations. We use three types of evaluation: comparison of extracted ***concept*** *relation* **feature** labels with a modified version of the McRae norms, a comprehensive direct human evaluation of our generated relations/features and a semantic similarity task, measuring the ability of our extracted labels to predict concept-concept similarity (using both human- and WordNet-derived similarity metrics).

3.1. Gold standard evaluation

In NLP, it is typical to evaluate performance by comparing system output with a gold standard. To do this, we will calculate precision, recall, and F-scores against a gold standard, such as (a subset of) the McRae norms. The precision score for a given concept's features is defined as the size of the overlap between the correct features for that concept (where a feature is defined as "correct" if it exists in the McRae norms) and the extracted features for that concept, divided by the total number of extracted features (i.e., the fraction of retrieved features which are relevant to the concept). Recall is defined as the size of the overlap between correct features and extracted features divided by the total number of correct features (i.e., the fraction of relevant features which are successfully retrieved). The F-score is the harmonic mean of precision and recall.

As has previously been discussed (Baroni, Evert, & Lenci, 2008; Kelly et al., 2010), employing these measures directly on the McRae norms presents additional difficulties in a true evaluation of system output to those discussed earlier, since lexically different statements often represent virtually identical features (e.g., ***jar*** *used for* **jelly** and ***jar*** *used for* **jam**). Therefore, we will employ the gold standard used at the ESSLLI Lexical Semantics Workshop 2008, Task 3: "Generation of salient properties for concepts" (Baroni et al., 2008).

Baroni et al. created a new gold standard which comprised the top 10 features for each of 44 concepts from the recoded McRae norms (the concepts belonged to six semantic categories: four animate and two inanimate). In addition to the 10 features for each concept, an "expansion set" was generated for each concept-feature pair to undo the normalization process described in our *Property Norms* section. This expansion set was constructed by first extracting from WordNet the synonyms of the words that constituted the concepts' features, then manually filtering out irrelevant synonyms and finally inserting other potential matches, including inflectional and derivational variants (**leg** for **legs** and **transport** for **transportation**), as well as other semantic neighbors or closely related entities. For example, **water** was expanded to **aquatic**, **lake**, **ocean**, **river**, **sea**, and **water**. This expansion relied on somewhat subjective human judgments. However, we, like the workshop authors, believe it offers a better evaluation than comparing directly against the McRae norms, as it allows comparison of extracted feature labels to the gold standard without insisting on exact lexical matching. Employing this set also allows us to directly compare our results with those of Kelly et al. (2010).

We note, however, that this set does not include expansions of relation labels. McRae observed that in the responses offered for the norming study "there was a great deal of variance," and the final choices of relations found in the McRae norms was the result of number of contributing factors: Some were plainly obvious, while others had historical precedent (e.g., *a* for the *is a* relationship) or had been chosen to illustrate certain semantic distinctions (e.g., ***car*** *has* **wheels** vs. ***car*** *requires* **driver** rather than ***car*** *has* **driver** since **driver** is not a part of a car).

Furthermore, there are often multiple ways to express the relationship between a concept and its feature. For example, consider the phrases "cars have doors," "the car doors," "she opened the door and got into her car," "this car's a three-door." We are hoping to funnel all of these expressions into our ***concept*** *relation* **feature** pattern, yet we can see that the *relation* portion (in this case *has* of ***car*** *has* **doors**) may not explicitly appear in the sentences we are extracting from. There is also little semantic difference between, for example, ***car*** *has* **doors**, ***car*** *includes* **doors**, yet we will only consider one of these to be correct when comparing directly against the McRae norms. We do not have a synonym-expanded set of relation labels to compensate for the above issues and therefore when comparing our extracted relations directly with the norms, we should bear in mind that this constitutes a reasonably tough standard of evaluation.

In our experiments we choose, in contrast to Baroni et al., to optimize and evaluate our system based on the top 20 extracted triples. We do this because it is simply not the case that all concepts possess only 10 properties (the majority of the McRae properties have more, and we have already discussed the incompleteness of the McRae norms). Although this will automatically will lower our highest possible scores (the average McRae concept has 14.7 features), we believe that including the top 20 offers a better insight into actual performance (all the more so when later performing our human evaluation). Doing this also offers us a better picture of how our reweighting stage is affecting our results, since we are considering a larger sample of our highly ranked triples. We will also examine performance for the top 10 extracted triples to offer like-for-like comparison with the evaluation of Baroni et al.

Since we are taking the top 20 triples, our results are not directly comparable with those of Devereux et al. either, who evaluated their system on the top 25% features returned. Throughout this section, we compare our results with the best-performing method from a preliminary version of our system (as described in Kelly et al., 2010) as this system has been evaluated identically to our own. This preliminary system has a similar structure to our new system, but it uses the RASP parser rather than C&C, has a different, more permissive candidate triple-extraction rule-set (which makes only one pass over the corpus), and reweights its extracted triples using the Semantic Reweighting Factor alone. It was also both optimized for and evaluated on the ESSLLI set, making it a good "best-possible" score to aim for.

3.2. Training

To avoid overfitting our system to a specific set of concepts, we will train it (and the free variables described at the end of the previous section) on a subset of our concepts. Then, we will evaluate our system on an unseen set of concepts. As already mentioned, we intend to use the 44 ESSLLI concepts (with relation expansion labels) for our final evaluation, and therefore we employ the remaining 489 McRae concepts to train our system. We do not have an expansion set for these triples, and so we must train and evaluate using exact matching on those relations and features found in the McRae norms.

We begin by examining how well our system is performing when applied to the training data prior to the reweighting stage (i.e., when all the β values are 0). These results can be found in Table 3. It is clear that ranking the triples by frequency alone does not offer particularly strong results.

Table 3
Precision, recall, and F-scores for all extracted top 20 triples (ranked by frequency) when evaluating against the non-ESSLLI norms, both including and excluding the relation

	Corpus	Prec	Rec	F
With relation	Wiki500	0.0307	0.0455	0.0358
	Wiki100K	0.0290	0.0449	0.0346
	BNC	0.0270	0.0403	0.0318
	Wiki100K-BNC	0.0348	0.0537	0.0415
Without relation	Wiki500	0.0909	0.1295	0.1033
	Wiki100K	0.0870	0.1332	0.1035
	BNC	0.1060	0.1590	0.1235
	Wiki100K-BNC	0.1141	0.1764	0.1363

3.2.1. Combining corpora

A quantitative analysis of the output indicates that there is actually relatively small overlap in the output of the two larger corpora: There is an overlap of 27.0% between the output of the Wikipedia set and the output of the BNC set, and an even smaller overlap of 14.6% when also taking the relation into account. It therefore seems worthwhile to evaluate the extent to which the triples from both corpora complement one another by measuring the performance when evaluating on a combination of the two (i.e., combining the output triples and scores from both sets and retaining the top twenty scoring triples from this combined set). The precision and recall results for this extraction set, which we call Wiki100K-BNC, can also be found in Table 3. The combination of these two larger corpora offers a slight improvement on our best scores (from the BNC corpus alone), indicating that this is indeed a viable approach.

3.2.2. Parameter estimation

The final stage of training optimizes our triple reranking schema for superior F-scores against our training set of concepts (i.e., the non-ESSLLI concepts). We vary our β parameters (described in the *Reweighting metrics* section) in the range [0,1] with an initial increment of 0.01, and then used the best-performing values to search for local F-score maxima around these values with increments of 0.001. We report our best training scores (with corresponding parameters) in Table 4. The most significant improvements from the reweighting appear in the combined corpus, where the F-score increases from 0.1363 to 0.1596. It is also interesting to note that the reweighting has not affected the results of the BNC output; although the triple output has changed due to the reweighting, the number of correct triples has not.

In general, we can see the reweighting favors the SRF metric; however, the entropy parameter also figures across seven of our eight different optimized systems, indicating that this is indeed a feasible metric. This might be somewhat surprising, especially considering how few rules we are employing in the extraction stage. The PMI and LL weightings are more variable (and less important) in their contributions to the final system, PMI notably not helping at all when reweighting triples with their relation included.

Table 4
Precision, recall, and F-scores for our system with corresponding β values when optimized against the non-ESSLLI norms, both ignoring the relation and including the relation

	Corpus	Prec	Rec	F	βLL	βPMI	βEnt	βSRF
With relation	Wiki500	0.0330	0.0492	0.0386	0.04	0.00	0.00	1.00
	Wiki100K	0.0414	0.0625	0.0490	0.04	0.00	0.02	0.98
	BNC	0.0367	0.0543	0.0430	0.00	0.00	0.04	0.9
	Wiki100K-BNC	0.0502	0.0764	0.0596	0.00	0.00	0.05	0.68
Without relation	Wiki500	0.094	0.1346	0.1071	0.00	0.05	0.03	1.00
	Wiki100K	0.1065	0.1622	0.1265	0.03	0.00	0.04	1.00
	BNC	0.1197	0.1787	0.1394	0.00	0.02	0.14	0.76
	Wiki100K-BNC	0.1339	0.2057	0.1596	0.03	0.03	0.09	0.78

3.3. Evaluation

3.3.1. Pre-reweighting results

Having estimated and fixed our parameters based on the non-ESSLLI concepts, we are now able to evaluate our system using those parameters. As already mentioned, we are evaluating on the top 20 returned triples. We note that when comparing with the ESSLLI gold standard, we are actually incorporating an upper bound for precision of 0.500 as the ESSLLI set contains only 10 properties per concept. We do this because we are aware that the gold standard is incomplete, and therefore it is plausible that we are extracting triples which are indeed correct but will be evaluated as wrong when compared with the ESSLLI set—we will discuss this in more detail in the sections that follow.

Although we are aiming to evaluate our system in its entirety (i.e., when considering our post-extraction statistics), it is illustrative to see how well the initial rule-based extraction method is performing on our 44 ESSLLI concepts. In Table 5, we report the results for all of our system output (ignoring the relation terms). Although precision is still low (between 1% and 4% for our new method), this is because there is no filtering on our output and there are thousands of triples being evaluated for each corpus (42,777 triples for Wiki500, 515,228 for Wiki100K, and 568,793 for the BNC corpus); this output is extremely voluminous compared to the number of "correct" triples (440), placing an extremely low upper bound on precision. This is also why the Wiki500 corpus tends to perform better—the corpus is much smaller than the other two and therefore produces fewer triples. These results would appear to indicate that in the initial feature extraction, we have increased our precision by a notable margin without an enormous loss of recall, especially for the Wiki100K and BNC corpora, when comparing to our preliminary system. The change is less impressive for the Wiki500 corpus with quite a strong reduction in recall, but this might be explained by the fact that our method is much more restrictive than our preliminary system in its candidate feature extraction. As such, the reasonably good results from the preliminary system on the Wiki500 corpus are a consequence of this relatively small and task-specific corpus being more likely to contain correct properties and less noise. It is also the corpus we employed when developing our set of rules.

Table 5
Precision, recall, and F-scores for all extracted triples, pre-reweighting, when evaluating against the ESSLLI norms, both including and excluding the relation term

		Corpus	Prec	Rec	F
Kelly et al. system (pre-reweighting)	With relation	Wiki500	0.0242	0.6515	0.0467
		Wiki100K	0.0039	0.8944	0.0077
		BNC	0.0042	0.8813	0.0083
	Without relation	Wiki500	0.1159	0.2326	0.1547
		Wiki100K	0.0761	0.1523	0.1015
		BNC	0.0841	0.1692	0.1123
Our system (pre-reweighting)	With relation	Wiki500	0.0312	0.0614	0.0413
		Wiki100K	0.0318	0.0636	0.0424
		BNC	0.0341	0.0682	0.0455
		Wiki100K-BNC	0.0375	0.0750	0.0500
	Without relation	Wiki500	0.0924	0.1818	0.1223
		Wiki100K	0.1000	0.2000	0.1333
		BNC	0.1420	0.2841	0.1894
		Wiki100K-BNC	0.1341	0.2682	0.1788

3.3.2. *Matching on features only*

In Table 5, we report our results when matching on the top 20 features only, ordered by frequency and prior to any reweighting. Here, we can again see an improvement almost entirely across the board; our new extraction system is clearly in the first instance generating more sensible triples than that of the Kelly et al. system. However, there is a slight reduction in F-score for the Wiki500 corpus. We feel this is again due to our more restrictive rule-set, which has prevented large numbers of less-than-certain features from being extracted. However, again since the Wiki500 corpus is smaller and more task-specific, these "less-than-certain" features are in fact more likely to be relevant, which might explain the higher F-score displayed by our preliminary system.

3.3.3. *Matching on features and relations*

We are not solely focusing on the features extracted; we also want to evaluate the relations. As already mentioned, matching on paired features and relations is a much more challenging task than matching on features alone, made all the more difficult by the fact that we do not have a synonym-expanded set of relations to evaluate on—we are evaluating directly on the relations found in the McRae norms. Our results can be found in Table 6. These results correspond to our final system.

Our final system is not performing quite as well as our "best-possible" method from Kelly et al. (2010). We believe this is for a number of reasons. For one, the "best-possible" system did not include a blind training phase. Instead, the system was optimized directly against the evaluation set, both by varying cluster size and clustering technique, to yield "best-possible" results against the ESSLI gold standard. Our system is also more conservative in the relations it is extracting, and more likely to only extract relations which it is confident in. Furthermore, as the ESSLLI expansion set we are comparing with does not include synonyms for the relation verbs, it is possible that the final step of our first stage

Table 6
Precision, recall, and F-scores for all extracted triples on our post-reweighting, final system when evaluating against the ESSLLI norms, both including and excluding the relation term. We compare our results with the Kelly et al. "best-possible" system

		Corpus	Prec	Rec	F
Kelly et al. system	With relation	Wiki500	0.1011	0.2028	0.1349
		Wiki100K	0.1102	0.2210	0.1471
		BNC	0.0955	0.1917	0.1275
	Without relation	Wiki500	0.1693	0.2326	0.1960
		Wiki100K	0.1733	0.3533	0.2325
		BNC	0.1943	0.3896	0.2593
Final system.	With relation	Wiki500	0.0323	0.0636	0.0428
		Wiki100K	0.0432	0.0864	0.0576
		BNC	0.0557	0.1114	0.0742
		Wiki100K-BNC	0.0602	0.1205	0.0803
	Without relation	Wiki500	0.1015	0.2000	0.1344
		Wiki100K	0.1227	0.2455	0.1636
		BNC	0.1420	0.2841	0.1894
		Wiki100K-BNC	0.1489	0.2977	0.1985

(in which we attempt to group similar triples) may be backfiring, in that although it is upweighting correct features (as demonstrated by our earlier results), it may be retaining an "incorrect" relation, and thus not performing as well as our benchmark system (which did not incorporate such a step). For example, ***helicopter*** *have* **pilot** is the highest rated triple with **pilot** as a feature in our system, and hence subsumes all instances of the correct triple from the ESSLLI set, ***helicopter*** *require* **pilot**. In other words, as our system retains only the highest-scoring relation when grouping triples with differing relations but the same features, exactly emulating the specific McRae-derived relations is challenging. This again demonstrates the pitfalls associated with this particular evaluation technique.

3.4. Human evaluation

We have already discussed many of the issues associated with employing the McRae/ESSLLI norms as our only point of comparison. We therefore turn to human evaluation: It is arguably the ultimate arbiter of whether the triples we are extracting are indeed correct. Although some properties may not be easily verbalizable or might not come to mind when people list properties during a property norming study, humans can still evaluate whether a given property is true with relative ease. That said, and as already mentioned, if a property truly is not verbalizable, it will be decidedly absent in any corpus we use.

From the 44 concepts appearing in the ESSLLI set, we chose a selection of 15 upon which to carry out our human evaluation. When selecting the concepts from the 44 candidate concepts, we first excluded three of them: ***snail*** (as it had only 9 features listed in the McRae set), ***onions*** (because it appeared in its plural form in the McRae set but as singular in the ESSLLI set), and ***truck*** (because this is known as "lorry" in British English, the dominant dialect of the BNC). The remaining 41 concepts had already been classified into ten superordinate categories (e.g., "animal") for unrelated psycholinguistic research, and we selected 15 concepts proportionally and at random from these superordinate categories. The selected concepts were: ***car***, ***cup***, ***duck***, ***hammer***, ***kettle***, ***knife***, ***lettuce***, ***lion***, ***motorcycle***, ***penguin***, ***pig***, ***pineapple***, ***potato***, ***screwdriver***, and ***turtle***.

Four native English-speaking judges evaluated the validity of both the extracted concept-relation-feature triples and concept-feature pairs.[6] They were asked to choose between four possibilities for each triple: *correct* (c) when the triple represented a correct, valid, property; *plausible* (p) when the triple was plausible in a very specific set of circumstances and/or was correct but very general; *wrong but related* (r) when the triple was wrong, but there existed some kind of relationship between the concept and the relation and/or feature; or *wrong* (w) when the triple was simply incorrect. The full criteria for our four possible categories for each triple differed slightly depending on whether the judge was evaluating a full concept-relation-feature triple or just a concept-feature pair (when performing the rating task on concept-feature pairs the relation was presented in the triple as an arrow). We include the full text of the two sets of instructions for the judges in the Appendices ("with relation" and "without relation" instructions appear in Appendix A and B, respectively).

Thus, the human evaluation was executed against our output both with and without relations, across our four corpora (our three initial corpora, plus the combined Wiki100K/BNC corpus) and across all 300 triples (15 concepts × 20 triples) for each corpus. As there were shared triples across this output, each distinct triple was evaluated only once. The judges were unaware of the purpose of the study, and the evaluation was done blind with regard to the source extraction set for each triple (thus making a deliberate bias toward any one of the extraction sets impossible).

Although we asked our annotators to allocate each of the triples to one of our four categories (*correct*, *plausible*, *wrong but related*, and *wrong*), we did this specifically to obtain more data for performing qualitative error-analysis of the system for future improvements, as well as to facilitate interpretation of the judgments themselves. We hope to use the human judgments of *correct* ("c") triples as training data for the next iteration of our system. However, given the subjective nature of the judgments for the purposes of measuring inter-annotator agreement, we consider all triples judged as *correct* or *plausible* merely to be *correct* (since, given the above definition of *plausible*, these triples are indeed correct, even if it is in a specific set of circumstances), and all those marked as *wrong but related* or *wrong* to be *incorrect*. We measure the degree of inter-annotator agreement based on whether a triple is judged to be *correct* or *incorrect* by our four judges. We use Fleiss' method (Fleiss, 1971) to calculate Kappa scores (Cohen, 1960) from our four annotators. Detailed agreement results when including the relation, can be seen in Table 7, with a highest Kappa score of 0.427 ("moderate" agreement according to the labels assigned to various Kappa ranges by Landis and Koch (1977)) for the Wiki100K corpus. The corresponding set of results when excluding the relation terms can be found in Table 8. Here, the Kappa scores are slightly higher across the corpora, which is perhaps somewhat surprising since there is less information for our human judges to base their decision on, and hence one might expect them to be more likely to disagree. Throughout we can see that on average at

Table 7
Triple ratings and inter-annotator agreement for our four corpora, evaluating the final system with the relation included. "Fully agree" means all four annotators gave the same rating (i.e., either c/p or r/w)

		Judge				
Corpus		A	B	C	D	Avg
Wiki500	c/p	226	154	194	197	192.75
	r/w	74	146	106	103	107.25
		150 full agreements, 50.0% fully agree				
	κ: 0.401	Correct/plausible: 64.25%				
Wiki100K	c/p	217	162	184	208	192.75
	r/w	83	138	116	92	107.25
		152 full agreements, 50.7% fully agree				
	κ: 0.427	Correct/plausible: 64.25%				
BNC	c/p	231	175	208	235	212.25
	r/w	69	125	92	65	87.75
		181 full agreements, 60.3% fully agree				
	κ: 0.361	Correct/plausible: 70.75%				
BNC-WIKI100K	c/p	237	185	208	229	214.75
	r/w	63	115	92	71	85.25
		168 full agreements, 56.0% fully agree				
	κ: 0.414	Correct/plausible: 71.58%				

Table 8
Triple ratings and inter-annotator agreement for our four corpora, evaluating the final system but excluding the relation. "Fully agree" means all four annotators gave the same rating (i.e., either c/p or r/w)

		Judge				
Corpus		A	B	C	D	Avg
Wiki500	c/p	222	219	175	201	204.25
	r/w	78	81	125	99	95.75
		168 full agreements, 56.0% fully agree				
	κ: 0.444	Correct/plausible: 68.08%				
Wiki100K	c/p	226	233	192	222	218.25
	r/w	74	67	108	78	81.75
		188 full agreements, 62.7% fully agree				
	κ: 0.486	Correct/plausible: 72.75%				
BNC	c/p	208	206	195	201	202.5
	r/w	92	94	105	99	97.5
		194 full agreements, 64.7% fully agree				
	κ: 0.572	Correct/plausible: 67.50%				
BNC-WIKI100K	c/p	232	236	217	235	230
	r/w	68	64	83	65	70
		207 full agreements, 69.0% fully agree				
	κ: 0.531	Correct/plausible: 76.67%				

Table 9
Judgments for the ordered top 20 triples for two concepts from our final system output. A "✓" indicates that the triple is correct according to the ESSLLI evaluation set

	Judge					Judge			
Triple	A	B	C	D	Triple	A	B	C	D
car *can be* **motor**	c	c	c	c	***penguin*** *can be* **king**	c	c	w	c
car *can be* **sport**	c	c	c	p	***penguin*** *be* **mascot**	c	p	p	p
car *have* **crash**	c	c	p	c	***penguin*** *be* **species**	c	c	c	c
car *have* **park**	r	p	c	c	***penguin*** *be* **game**	p	p	w	p
car *have* **accident**	c	c	p	c	***penguin*** *can be* **adelie**	c	w	w	w
car *can be* **electric**	c	c	c	c	***penguin*** *be* **character**	p	p	p	c
car *be* **vehicle** (✓)	c	c	c	c	***penguin*** *can be* **young**	c	c	c	c
car *have* **door** (✓)	c	c	c	c	***penguin*** *can be* **emperor**	c	c	c	c
car *can be* **passenger**	c	c	c	c	***penguin*** *have* **book**	c	r	p	c
car *do* **drive**	p	c	c	c	***penguin*** *can be* **african**	w	c	w	w
car *do* **run**	c	c	p	c	***penguin*** *be* **large**	c	p	c	p
car *have* **parking**	p	p	c	c	***penguin*** *be* **book**	c	r	p	c
car *can be* **racing**	c	c	c	p	***penguin*** *be* **hoax**	p	w	w	p
car *have* **driver**	c	c	c	c	***penguin*** *can be* **male**	c	c	c	c
car *can be* **private**	c	c	c	c	***penguin*** *be* **adelie**	p	c	r	w
car *can be* **small**	c	c	c	c	***penguin*** *be* **tall**	c	p	c	p
car *have* **engine** (✓)	c	c	c	c	***penguin*** *can be* **giant**	c	c	c	r
car *can be* **fast**	c	c	c	c	***penguin*** *be* **seal**	p	r	r	r
car *have* **crime**	r	w	p	c	***penguin*** *name* **humboldt**	c	r	w	r
car *have* **racing**	c	r	c	c	***penguin*** *do* **fly**	w	p	w	w

least half of those triples marked as incorrect by our ESSLLI evaluation (both including and excluding the relation) are considered *correct* or *plausible* by our judges. The judgments given for two concepts can be found in Table 9.

Examining the output we can see that for ***car***, the vast majority of our output is deemed correct by humans; the cases where there is disagreement often point to a feature which is subjectively linked to the concept at hand (e.g., ***car*** *have* **crime**), but on the whole the features extracted are indisputably associated in some way with the concept at hand. For the concept ***penguin***, there is more disagreement between the annotators; this may be because some rather technical terms have been extracted from the corpus (our system has extracted the names of five separate species of penguin) and it is a subjective question as to the extent to which these species are "features" of the concept ***penguin***. Furthermore, the fact that not one of the extracted features (many of which have been deemed correct or plausible by our human judges) appear in the ESSLLI evaluation gold standard again demonstrates the very difficult nature of that particular evaluation task.

3.4.1. Combining McRae and human evaluation

Given that we have collected our human-evaluation data, it might also be instructive to assess—using the human ratings—how the system is performing for our 15 concepts. To do this, we calculate the precision for the top 20 features for each of the concepts from the sets (we are unable to calculate recall and F-scores because there is no upper bound on the number of triples which our human judges could deem correct). Each triple is marked as correct if and only if it is marked as either plausible or correct by *all* our judges. Our results are shown in Table 10. In this table, we also include, for comparison, the precision scores for the same set of triples when using the ESSLLI evaluation methodology.

Since this evaluation is across only a relatively small number of concepts, it should not be interpreted as the full picture of how our system is performing. However, it does signal the extent to which patently incorrect triples are appearing. Our results indicate that just under half of the triples returned are correct, and when evaluating on features alone, it is possible to achieve precision scores of over 60%. This exceeds the corresponding best ESSLLI precision score of 16% by far, further demonstrating the limitations of the direct evaluation technique.

3.5. Human-generated semantic similarity comparison

Given the issues associated with calculating precision and recall scores directly from our output, we use an additional, alternative approach to calculate how semantically meaningful our extracted triples really are. To do this, we evaluate their capacity to predict similarity between words, using human similarity judgments.

We asked 10 native English speakers to rate the similarity of 90 concept pairs. The concepts were all drawn from the ESSLLI set. The 90 concept pairs correspond to 10 concept pairs chosen at random from their banded WordNet similarity score, based on Leacock and Chodorow's Normalized Path Length WordNet similarity measure (Leacock and Chodorow, 1998). In other words, there were 10 concepts with score 0–0.1, 10 with score 0.1–0.2, and so on. There were no pairs with similarity of 0.9 or above.

The raters were given the following instructions:

You will be presented with pairs of words that refer to well-known concepts in the world (e.g., "turtle ⟨–⟩ kettle," "lion ⟨–⟩ dog"). Your task is to rate, on a scale of 1 to 7, how similar in meaning the two concepts are. They were then presented with each con-

Table 10
Precision scores for our final system when evaluating against the human judgments on 15 concepts both with features only (FO) and with features and relations (F&R). For comparison, we also report precision scores for the same set of output triples using the ESSLLI evaluation methodology

	Human		ESSLLI	
Corpus	FO	F&R	FO	F&R
Wiki500	0.4500	0.3733	0.1000	0.0333
Wiki100K	0.5233	0.3867	0.1233	0.0367
BNC	0.4900	0.4967	0.1400	0.0633
Wiki100K-BNC	0.5967	0.4767	0.1600	0.0666

cept pair, one by one, a scale of 1 to 7 and asked: *Please rate how similar the following concrete nouns are on a scale of 1 to 7, 1 meaning "very dissimilar," and 7 meaning "very similar."*

The average Pearson coefficient of correlation across the 10 judges (considering all pairwise combinations) was 0.8204.

Using these scores, we constructed a vector of dimensionality 90, *VHuman* containing the averaged human-generated similarity scores between the 90 concept pairs. We normalized each score so that it lay between 0 and 1 (i.e., "very dissimilar" pairs received a score of 0, "very similar" pairs received 1, and the remaining scores were distributed evenly across the interval).

To compare our system with these ratings, we wish to approximate the similarity between our 44 ESSLLI concept words given a set of m output triples for each concept (and from this we will be able to extract similarity vectors V corresponding to the 90 pairwise human comparisons). To achieve this, we begin by constructing a vector space of dimension D, where D is the number of distinct properties across our $44 \times m$ triples. Then for each of our 44 concepts, we generate a concept-score vector with 20 non-zero entries by inserting the triple scores, score(t), into their correct entries in the concept-score vector. We may then construct a 44×44 symmetric pairwise similarity matrix across our 44 concepts by calculating the cosine similarity between their concept-score vectors. From this we can extract similarity vectors, V, for our 90 concept pairs.

We perform our calculations for $m = 20$, corresponding to our optimization for and previous evaluations of the top twenty triples thus far. However, one of the benefits of our system is its ability to generate a very large number of properties for each concept, and therefore we also evaluate the correlations for $m = 300$, thereby taking a large number of the extracted triples into account. For each value of m, we calculate eight such matrices: both including and excluding the relation term from each concept's triple across our four corpora.

We similarly generate two similar matrices from the McRae norms (one using the full text of the property norms as the concept vectors' dimensions, the other using only the feature heads), using the norm production frequencies (in place of score(t)) as entries in each concept's vector.

Finally, for comparison we also correlate the human ratings with a similarity vector derived from Latent Semantic Analysis (Deerwester et al., 1990) applied to the "General Reading up to First Year College" corpus with $k = 300$ factors.[7] Selection of an optimal number, k, of LSA factors is known to be challenging; however, correlations between human term-to-term similarity judgments and LSA similarity values are often highest for values of $k \approx 300$ (Dumais, 1994; Steyvers, Shiffrin, & Nelson, 2004). Like our human-generated ratings, the cosine similarity values were normalized to lie between 0 and 1.

In this way, we may report the Pearson correlations between our *VHuman* vector and our various similarity vectors V. Our results can be found in Table 11. We first note that the similarity values derived from the McRae norms exhibit the strongest correlation with the human-generated ratings, and the inclusion/exclusion of the relation term in these norms does not appear to have a significant impact on the results (correlation of around

0.786 for both). The LSA Pearson correlation is also strong at 0.7076. However, when considering our system's feature-only top 20 output on the BNC corpus, we achieve 0.6655 correlation, which falls well within the LSA correlation confidence interval. Furthermore, when considering a larger sample of our extracted triples (the top 300) from the combined corpus, our system's similarity values are capable of exceeding LSA's semantic similarity predictive abilities, reaching a Pearson correlation with human respondents of 0.7411.

In light of these results, we offer a brief examination of the lower ranked output to illustrate some triples which have contributed to our improved correlation scores. Table 12 shows the 101st to 110th and 201st to 210th top triples as output by our final system for two concepts, as well as four judges' ratings for these triples. It is clear that a majority of the lower-ranked triples are still—despite their lower ranking—related or relevant to the concept at hand. Having a large number of such (related) properties creates a more refined representation of the concept at hand, and by increasing the potential for properties to be shared across concepts, it facilitates measurement of semantic similarity by way of the "sharedness" of two concepts' properties.

3.6. WordNet semantic similarity comparison

Finally, we also compare our output with the semantic similarity predicted by WordNet across all pairwise combinations of our 44 concepts. To achieve this, we construct a baseline matrix, *MLC* containing similarity scores between all binary combinations using the Leacock and Chodorow WordNet similarity measure. We normalized the returned values by dividing through by the maximum possible value, ensuring that all values lay in the range [0,1].

Table 11
Pearson correlation results between our *VHuman* vector and our similarity vectors *V* from our final system as reported in Table 6. The confidence intervals, calculated using Fisher transformations, are given at the 95% level of confidence, and two-tailed $p < .05$ for all our correlation calculations. We report results when considering features only (FO) and when evaluating on features and relations (F&R)

		r		Confidence Interval	
V		FO	F&R	FO	F&R
McRae		0.7874	0.7853	[0.693, 0.855]	[0.691, 0.854]
LSA (300 factors)		0.7076		[0.586, 0.798]	
Top 20 output	Wiki500	0.4684	0.3194	[0.289, 0.616]	[0.120, 0.494]
	Wiki100K	0.4897	0.3927	[0.314, 0.633]	[0.202, 0.555]
	BNC	0.6655	0.5625	[0.532, 0.767]	[0.402, 0.689]
	Wiki100K-BNC	0.6305	0.5452	[0.487, 0.741]	[0.381, 0.676]
Top 300 output	Wiki500	0.4738	0.3217	[0.296, 0.620]	[0.123, 0.496]
	Wiki100K	0.5880	0.4698	[0.434, 0.709]	[0.291, 0.617]
	BNC	0.7037	0.5953	[0.581, 0.795]	[0.443, 0.714]
	Wiki100K-BNC	0.7411	0.6440	[0.631, 0.822]	[0.504, 0.751]

Table 12
Judgments for the 101st–110th and 201st–210th extracted properties from our final system output for two concepts

	Judge					Judge			
Triple	A	B	C	D	Triple	A	B	C	D
101st–110th									
car *can be* **kit**	p	r	p	c	***penguin*** *be* **cap**	r	p	w	w
car *be* **good**	P	c	r	c	***penguin*** *do* **sit**	p	c	p	c
car *have* **company**	c	c	r	c	***penguin*** *can be* **crested**	p	c	p	c
car *have* **showroom**	p	c	c	c	***penguin*** *be* **flightless**	c	c	p	c
car *can be* **company**	c	p	p	w	***penguin*** *be* **equivalent**	r	p	w	w
car *have* **buyer**	c	c	c	c	***penguin*** *be* **depiction**	p	p	w	w
car *have* **sales**	c	c	c	c	***penguin*** *be* **attraction**	c	c	p	w
car *do* **race**	p	p	c	c	***penguin*** *adapt* **life**	c	c	p	c
car *can be* **parked**	c	c	c	c	***penguin*** *have* **island**	p	c	p	c
car *be* **subject**	r	c	p	w	***penguin*** *can be* **related**	c	c	c	c
201st–210th									
car *can be* **patrol**	c	p	c	p	***penguin*** *have* **brand**	c	p	r	w
car *have* **market**	p	c	c	p	***penguin*** *follow* **character**	r	r	r	w
car *be* **bus**	r	r	r	r	***penguin*** *first* **disguise**	r	r	r	w
car *can be* **found**	p	c	p	r	***penguin*** *feature* **character**	p	p	p	w
car *can be* **formula**	W	p	r	w	***penguin*** *feature* **album**	w	p	w	w
car *can be* **many**	c	c	p	r	***penguin*** *endemic* **antarctica**	c	r	c	p
car *have* **model**	c	c	c	c	***penguin*** *endanger* **pair**	c	c	p	p
car *can be* **freight**	w	r	p	p	***penguin*** *do* **wander**	c	c	c	c
car *can be* **blue**	p	c	c	c	***penguin*** *do* **voice**	c	c	w	c
car *be* **falcon**	w	p	w	w	***penguin*** *do* **talk**	r	p	w	r

The Frobenious norm of a matrix X is defined as follows:

$$||X||F = \sqrt{\sum_{i,j} |x_{ij}|^2} \tag{7}$$

We can calculate the Frobenious distance between two matrices X and Y as $||X - Y||F$. A lower Frobenious distance between two similarity matrices implies that they are closer to one another, as it is the matrix equivalent of calculating the Euclidean distance between two points. Our results for measuring the Frobenious distances between our various matrices and *MLC* are shown in Table 13.

We also calculate the Pearson correlation between each of our matrices and the Leacock and Chodorow similarity matrix. Because our matrices are symmetric, we only want to take each pairwise similarity into account once, and we ignore the trivial identity similarities. Hence, we use the upper triangular versions of the matrices. Each upper triangular matrix U has $N = 43 \times 44/2$ entries above the main diagonal, corresponding to the 946 pairwise similarity values across all 44 words. We then calculate the Pearson correlation, r, between each of our matrices, U, and our baseline matrix *ULC*. We use the Fisher

Table 13
Frobenious distances between our similarity matrices and our baseline matrix based on the Leacock and Chodorow WordNet similarity metric. Lower values of $F = ||MLC - M||F$ indicate a closer relationship between the matrix in question and our baseline matrix, *MLC*. *D* values correspond to the matrices' dimensionalities. We also calculate $r = corr(ULC, U)$, the Pearson correlation after applying an upper triangular matrix transformation to our matrices. The confidence intervals, calculated using Fisher transformations, are given at the 95% level of confidence, and two-tailed $p < .05$ for all our correlation calculations. We report results when considering features only (FO) and when evaluating on features and relations (F&R)

	D		*F*		*r*		Confidence Interval	
M	FO	F&R	FO	F&R	FO	F&R	FO	F & R
McRae	355	410	15.32	15.53	0.4780	0.4721	[0.473, 0.483]	[0.467, 0.477]
Wiki500	626	712	16.76	17.22	0.2438	0.1553	[0.238, 0.250]	[0.149, 0.162]
Wiki100K	533	626	15.77	16.46	0.2989	0.2084	[0.293, 0.305]	[0.202, 0.215]
BNC	542	601	16.14	16.44	0.4170	0.3013	[0.412, 0.422]	[0.296, 0.307]
Wiki100K-BNC	524	586	16.13	16.26	0.3109	0.2568	[0.305, 0.317]	[0.251, 0.263]

transformation (Fisher, 1915) to report confidence intervals at the 95% level of confidence (Table 13).

The matrices derived from our BNC triples appear to be the best predictor of concept-concept similarity (both when including and excluding the relation terms), showing the highest overall correlations with the Leacock and Chodorow baseline similarity matrix and human evaluations. It is noteworthy that the correlations are higher for the features-only matrices across the board. This could be explained by the relatively low number of "shared" properties between concepts when including the relation and further exacerbated by the fact we are only using 20 properties to describe each concept. In general, the correlation scores against the WordNet baseline are lower than the corresponding scores against human-judged similarities in the previous section.

3.7. Conclusion

Over each evaluation method, our results indicate that we have taken a significant step in the right direction—notwithstanding the various evaluation issues which we have already discussed, our results indicate solid performance when compared to the "best-possible" system by Kelly et al. We have shown a high level of accuracy on our output when judged by humans: almost 60% precision when judging on features alone. Furthermore, our semantic vector comparison indicates that our final system is not trailing all that far behind the McRae norms themselves when predicting human-judged semantic similarity of concepts, and indeed is capable of outperforming LSA.

4. Discussion

This article has presented a state-of-the-art system capable of extracting verifiably plausible property norm-like features from large corpora. When evaluating on features alone our system can achieve human-judged precision scores of over 60%; human evaluation also showed that over 71% of the top 20 features with relations returned were "correct" or "plausible." The system is also capable of predicting human ratings of semantic similarity, showing a correlation of 0.7411 with human-judged similarities when using the top 300 output triples (compared to LSA's 0.7076 using 300 factors).

When comparing our system with that of Baroni et al. (2010), using their evaluation criteria (i.e., calculating precision and recall on the top 10 features only), our system (using the combined BNC/Wikipedia corpus and the reweighting parameters listed in Table 4) produces a best F-score of 0.207—their best F-score is 0.239—which we believe is a strong result, given the propensity for the evaluation to yield false negatives, and the fact that our system was optimized for the top 20 features rather than the top 10. It is also important to note that Baroni et al.'s method falls short of explicitly listing the relationships between the concepts and features it extracts; our system is ambitious in attempting to do so. We therefore believe our work forms an important first step toward accomplishing this highly challenging task.

In our work, we have demonstrated that the use of directional grammatical relation patterns and part-of-speech data derived from parsed corpus-data can be beneficial to extracting candidate concept-relation-feature triples. We generated rules over GR patterns manually, but future work could take a more generalized approach to this—for example, by examining the possibility of automating the rule-learning process in the first stage of our system to remove its manual element. To achieve this, we could employ semi-supervised training techniques to acquire the rules; such machine learning techniques have offered state-of-the-art performance for many NLP tasks. We could follow a similar approach to that taken by Mintz, Bills, Snow, and Jurafsky (2009), who created a relation classifier using a paradigm called "distant supervision" which assumed that "if two entities participate in a relation, any sentence that contains those two entities might express that relation." Such a system would differ from Mintz et al.'s in that it could use any combination of lexical and syntactic attributes, implicit and explicit, obtained from the path linking the concept to the feature to generate a rule. An optimal set of attribute-based patterns would need to be derived empirically: We could use a portion of known property norms as a training set to teach the system which patterns of GR-POS graph paths typically indicate plausible properties/triples.

Our work has also examined the relative capacity of four distinct metrics to upweight human-like features/relations. Our results indicate that two of these (the entropy of a triple, calculated from the probability mass across rules which generate it, and the semantic reweighting factor) offer improvements when evaluating against features alone and features with their relations. Our other two measures (pointwise mutual information and log-likelihood ratio) offer less marked improvements; however, both do contribute to certain 'best' systems, depending on the corpus employed. This could be explained by the fact that both of these measures will tend to prioritize more distinctive triples, but as the required degree of distinctiveness for accurate properties is likely to be concept-dependent the reweighting scheme tends toward extracting "safer," less idiomatic triples.

In future work, we could investigate further reweighting factors likely to yield human-like norms: for example, the *t*-test of word-association by Manning and Schütze (1999). Our current system is able to extract features reasonably well, but extracting correct relations seems to present more difficulties. This indicates that it might be worthwhile restructuring the system so that it first evaluates likely features using the various reweighting factors, only later returning to the corpus to find likely relations for those features.

This article has also shown the benefits that may be derived from combining multiple types of corpora: a simple concatenation of extracted triples from our two corpora offered an immediate improvement across the board. NLP techniques tend to perform better with more data, and therefore future work could employ further, larger corpora. One could also use simpler, more "basic" corpora; for example, a corpus of children's literature (e.g., Sealey and Thompson [2004] used a subset of the BNC containing only texts written for children) or a "basic English" corpus (e.g., Ruiz-Casado, Alfonseca, & Castells [2005] used Simple English Wikipedia to automatically extract semantic relations for WordNet). Our system tends to perform better with shorter sentences (as this renders the grammatical relation paths shorter and consequently less error-prone), and the type of "common-sense" data often found in such corpora could be of enormous benefit. We could also consider using an empirically derived, more sophisticated weighting of corpora to maximize accuracy—for example, certain corpora may be better for certain types of relations/features.

Improving our property representation is another potential avenue of research. In our work, we reduced our extracted features to a simplified three-part triple in the form ***concept*** *relation* **feature**. Future, more sophisticated representations could harness the flexibility in feature-derivation offered by our path-based rule construction system (i.e., the fact we can extract more than one node from within a matched path). These property differences are meaningful in terms of cognitive theories (e.g., Kremer & Baroni, 2010), especially so for distinctive properties such as ***giraffe*** *have* **long neck** which under our current representation would be recoded to ***giraffe*** *have* **neck**. They therefore motivate a flexible relation/feature representation. Similarly, introducing prepositions into our relation terms (e.g., ***anchor*** *find in* **water** instead of ***anchor*** *find* **water**) would be a significant step in terms of disambiguating their meaning (and would likely further improve our inter-annotator agreement scores).

As previous research (e.g., Devereux et al., 2009; Kelly et al., 2010) has recognized, finding an accurate and reliable evaluation methodology remains a serious obstacle in assessing the performance of any system approaching our task. We have employed the ESSLLI evaluation subset as our gold standard, but our human evaluation has demonstrated obvious deficiencies in it. To address these, our work also introduced a novel property evaluation method based on similarity matrices, which has shown our extracted triples to be capable of producing binary-concept similarity whose accuracy is not very far off that of the McRae norms when using Leacock and Chodorow's WordNet similarity metric as the baseline.

Finally, when predicting human similarity ratings, our system is capable of outperforming LSA. This is a noteworthy result—unlike LSA, our system was not explicitly designed to predict term-term similarity, rather to extract conceptual properties. Our system's correlation scores improved considerably when considering a larger number of properties (i.e., the top 300 extracted properties rather than the top 20), and this result—which was not all that far of the McRae norms' performance on the same task—was a product of this. We believe that as the McRae norms contain only a small number of accurate properties, there is potential for a computationally generated property-based semantic representation containing a large number of accurate properties to exceed the predictive ability of the McRae norms on such tasks.

We conclude our discussion by discussing the theoretical implications of our research. We begin by drawing attention to the fact that techniques for extracting properties of concepts—rather than mere word-associations or concept-similarity predictors—are still in their infancy. Furthermore, they must be viewed as distinct and, we hope, more useful than semantic classifiers/cluster generators, at least in the context of cognitive psychology.

We acknowledge that the properties we extract are not "behavioral" in the same way that McRae's norms are; rather they are, by their derivation, a product of our chosen rules and later by their statistical distribution in text. We are not contending that the properties are exactly equivalent to those found in people's minds or its equivalent representation (i.e., conceptual knowledge). Rather, we are aiming to demonstrate the breadth and richness of semantic information that is possible to extract from large bodies of text and that there is potential to generate a high proportion of the conceptual properties which humans know, with the additional benefit of using these properties for research in cognitive psychology.

An important criticism of property norming studies, from a cognitive psychology perspective, is that although they are interesting inasmuch as the presence/absence of certain properties is telling in itself, they are incomplete in terms of building a full and comprehensive property-based description of a certain concept. The fact that humans could most likely surmise the identity of a concept given only its properties from, for example, the McRae norms is a product of not only the cited properties but also participants' pre-existing conceptual knowledge which "fills in the gaps." In our view, an ideal system would generate all properties for a given concept, ranking them in terms of their specificity and salience for the concept at hand. Our work aims to fill in more of these gaps, but it is as yet unknown whether we can totally emulate all conceptual knowledge about a given concept. What is, however, clear is that a significant amount of semantic information—as rich as that found in property norms—can be gleaned purely from data in language. This may seem like a manifest statement, but it is an important one nonetheless. Our system's output is a product of the functional structure of language, rather than the structure of human knowledge or the world itself—this distinction is key if only because we still do not fully understand the nature of the differences/similarities between these categories. Yet it also rises to the question of whether, given appropriate NLP techniques and a sufficiently large corpus (which would effectively be the product of many human's thoughts and statements, derived both from their linguistic interactions and their experience of the world itself), it would be possible to entirely and comprehensively emulate the conceptual representation of a given concept by a layman. Can knowledge in language (particularly, with the advent of the web, the sum of many people's linguistic output) adequately mirror that in the mind or the world?

We leave it is an open question as to which technique cognitive psychologists can or should employ in attempting to understand how language is conceptualized in the brain: be it the sometimes-changing, often idiosyncratic (perhaps language/culture dependent) and highly concept-specific output of humans citing properties for a given concept; or a technique which is less intuitive by human standards but arguably offers a more scientific, consistent, and uniform approach to the representation of concepts.

The development and implementation of accurate property extraction methods and their evaluation is a challenging, and relatively new, task. Our work demonstrates that there is a wealth of semantic knowledge to be gained from language itself and reasonable results in extracting it are possible through the combination of a number of NLP techniques.

Notes

1. There is ambiguity in using **feature** to describe just the final term of such a triple, as the entire pattern (or sometime just the latter two terms) has in previous work been called a "feature." We adopt the convention that a "feature" is only ever a single term feature-head, **feature**, whereas a "property" describes the full property norm, in the form of a ***concept*** *relation* **feature** pattern.
2. However, using a corpus will not, for obvious reasons, yield properties which are not verbalizable.
3. For example, for ***car***, "used for transportation" and "people use it for transportation" were mapped to the same *used for* **transportation** relation/feature pair.
4. We chose these values so that the average cluster size would be around 10.
5. WordNet (Fellbaum, 1998) is a large lexical database of English, where words are grouped into synonym sets, each expressing a different concept.
6. The pairs and triples are collectively called "triples" in the paragraphs which follow.
7. Cosine similarity values between our 90 concept pairs were obtained using the publicly available Pairwise Comparison tool on the lsa.colorado.edu website.

References

Almuhareb, A., & Poesio, M. (2004). Attribute-based and value-based clustering: An evaluation. In D. Lin & D. Wu (Eds.), *Proceedings of EMNLP 2004* (pp. 158–165). Barcelona, Spain: Association for Computational Linguistics.

Almuhareb, A., & Poesio, M. (2005). Concept learning and categorization from the web. In B. G. Bara, L. Barsalou & M. Bucciarelli (Eds.), *Proceedings of CogSci* (pp.103–108). Stresa, Italy.

Andrews, M., & Vigliocco, G. (2010). The hidden Markov topic model: A probabilistic model of semantic representation. *Topics in Cognitive Science*, *2*(1), 101–113.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*(3), 463.

Barbu, E. (2008). Combining methods to learn feature-norm-like concept descriptions. In M. Baroni, S. Evert & A. Lenci (Eds.), *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics* (pp. 9–16). Hamburg, Germany: Association for Logic, Language and Information.

Baroni, M., Evert, S., & Lenci, A. (Eds.). (2008). *ESSLLI 2008 Workshop on Distributional Lexical Semantics*. Hamburg, Germany: Association for Logic, Language and Information.

Baroni, M., & Lenci, A. (2008). Concepts and properties in word spaces. *Italian Journal of Linguistics*, *20* (1), 55–88.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*(4), 673–721.

Baroni, M., Murphy, B., Barbu, E., & Poesio, M. (2010). Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, *34*, 222–254.

Bird, S. (2006). NLTK: The natural language toolkit. In J. Curran (Ed.), In *Proceedings of the COLING/ACL Interactive Presentation Sessions* (pp. 69–72). Sydney, Australia: Association for Computational Linguistics

Briscoe, T. (2006). An introduction to tag sequence grammars and the RASP system parser. *Computer Laboratory Technical Report*, 662.

Bruni, E., Tran, G., & Baroni, M. (2011). Distributional semantics from text and images. In S. Pado & Y. Piersman (Eds.), *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics* (pp. 22–32). Edinburgh, UK: Association for Computational Linguistics.

Burnard, L. (2007). Reference guide for the British National Corpus (XML Edition). URL: http://www.natcorp.ox.ac.uk/XMLedition/URG/

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Clark, S., & Curran, J. (2007a). Formalism-independent parser evaluation with CCG and DepBank. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Vol. *45*, p. 248. Sydney, Australia, Association for Computational Linguistics .

Clark, S., & Curran, J. (2007b). Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, *33*(4), 493–552.

Cohen, J. (1960). *A coefficient of agreement for nominal scales*. Vol. *20*, p. 37–46, Educational and Psychological Measurement.

Collins, M., & Singer, Y. (1999). Unsupervised models for named entity classification. In P. Fung & J. Zhou (Eds.), *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (pp. 100–110). Baltimore, MD, Association for Computational Linguistics.

Cree, G., McNorgan, C., & McRae, K. (2006). Distinctive features hold a privileged status in the computation of word meaning: Implications for theories of semantic memory. *Journal of Experimental Psychology Learning Memory and Cognition*, *32*(4), 643.

Davidov, D., & Rappoport, A. (2008). Classification of semantic relationships between nominals using pattern clusters. In Proceedings of ACL 2008, HLT, pp. 227–235.

Davidov, D., Rappoport, A., & Koppel, M. (2007). Fully unsupervised discovery of concept-specific relationships by web mining. In A. Zaenen & A. van den Bosch (Eds.), *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, *41*, 391–407.

Devereux, B., Pilkington, N., Poibeau, T., & Korhonen, A. (2009). Towards unrestricted, large-scale acquisition of feature-based conceptual representations from corpus data. *Research on Language & Computation*, *7*(2–4), pp 137–170.

Devlin, J., Gonnerman, L., Andersen, E., & Seidenberg, M. (1998). Category-specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of Cognitive Neuroscience*, *10*(1), 77–94.

Dumais, S. T. (1994). Latent semantic indexing (LSI) and TREC-2. In *The Second Text REtrieval Conference (TREC-2)*. Maryland, USA: National Institute of Standards and Technology (pp. 105–115).

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A., Shaked, T., Soderland, S., et al. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, *165*(1), 91–134.

Farah, M., & McClelland, J. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General*, *120*(4), 339–357.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Fisher, R. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*(4), 507–521.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378.

Garrard, P., Ralph, M., Hodges, J., & Patterson, K. (2001). Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, *18*(2), 125–174.

Grondin, R., Lupker, S., & McRae, K. (2009). Shared features dominate semantic richness effects for concrete concepts. *Journal of Memory and Language*, *60*(1), 1–19.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In A. Zampolli (Ed.), *Proceedings of the 14th International Conference on Computational Linguistics* Vol. *2* (pp. 539–545). Nantes, France: Association for Computational Linguistics.

Johns, B. T., & Jones, M. N. (2012). Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, *4*(1), 103–120.

Jurafsky, D., & Martin, J. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (Vol. *2*). London: Prentice Hall.

Kelly, C., Devereux, B., & Korhonen, A. (2010). Acquiring human-like feature-based conceptual representations from corpora. In *Proceedings of the First Workshop on Computational Neurolinguistics (p. 61)*. Los Angeles, CA: Association for Computational Linguistics.

Kremer, G., & Baroni, M. (2010). Predicting cognitively salient modifiers of the constitutive parts of concepts. In J. T. Hale (Ed.), *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics* (pp. 54–62). Uppsala, Sweden: The Association for Computational Linguistics.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33* (1), 159–174. Available from http://www.jstor.org/stable/2529310

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, *49* (2), 265–283.

Leech, G., Garside, R., & Bryant, M. (1994). CLAWS4: the tagging of the British National Corpus. In M. Nagao (Ed.), *Proceedings of the 15th International Conference on Computational Linguistics*: Vol *1* (pp. 622–628). Kyoto, Japan: International Conference on Computational Linguistics.

Lin, D. (1998). An information-theoretic definition of similarity. In J. W. Shavlik (Ed.), *Proceedings of the 15th International Conference on Machine Learning* (pp. 296–304). Morgan Kauffman.

Louwerse, M. M. (2010). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273–302.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. *59*). Cambridge, MA: MIT Press.

Masson, M. (1995). A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 3.

McRae, K., Cree, G., Seidenberg, M., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavioral Research Methods, Instruments, and Computers*, *37*, 547–559.

McRae, K., De Sa, V., & Seidenberg, M. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology-General*, *126*(2), 99–130.

Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In K-Y. Su, J. Su & J. Wiebe (Eds.), *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* Vol. *2* (pp. 1003–1011). Singapore: Association for Computational Linguistics.

Murphy, G. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Pantel, P., & Lin, D. (2002). Discovering word senses from text. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 613–619). Massachusetts, USA.

Pantel, P., & Pennacchiotti, M. (2008). Automatically harvesting and ontologizing semantic relations. In O. R. Zaïane (Ed.), *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge* (pp. 171–195). Amsterdam: IOS Press.

Poon, H., & Domingos, P. (2010). Unsupervised ontology induction from text. In J. Hajič (Ed.), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 296–305). Uppsala, Sweden: Association for Computational Linguistics.

Randall, B., Moss, H., Rodd, J., Greer, M., & Tyler, L. (2004). Distinctiveness and correlation in conceptual structure: Behavioral and computational studies. *Journal of Experimental Psychology Learning Memory and Cognition*, *30*(2), 393–406.

Rindflesch, T., Tanabe, L., Weinstein, J., & Hunter, L. (2000). EDGAR: extraction of drugs, genes and relations from the biomedical literature. In R. B. Altman, A. K. Dunker, L. Hunter & T. E. Klein (Eds.), *Pacific Symposium on Biocomputing*. New Jersey, USA, World Scientific. (p. 517).

Ruiz-Casado, M., Alfonseca, E., & Castells, P. (2005). Automatic extraction of semantic relationships for WordNet by means of pattern learning from Wikipedia. In A. Montoyo & E. Métais (Eds.), *Natural Language Processing and Information Systems*, *3513*, 67–79. Alicante, Spain.

Sealey, A., & Thompson, P. (2004). What do you call the dull words? Primary school children using corpus-based approaches to learn about language. *English in Education*, *38*(1), 80–91.

Steyvers, M. (2010). Combining feature norms and text data with topic models. *Acta Psychologica*, *133*(3), 234–243.

Steyvers, M., Shiffrin, R., & Nelson, D. (2004). *Word association spaces for predicting semantic similarity effects in episodic memory* (pp. 237–249). Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer.

Turney, P. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In L. De Raedt & P. Flach (Eds.), *Proceedings of the 12th European Conference on Machine Learning* (pp. 491–502). Freiburg, Germany, Springer-Verlag.

Tyler, L., Moss, H., Durrant-Peatfield, M., & Levy, J. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language*, *75*(2), 195–231.

Vinson, D., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, *40*(1), 183.

Appendix A

Human evaluation instructions: Part 1

In this experiment, you are asked to judge whether properties listed for concepts are true or not. The properties are listed as word triples of the form `<concept>` `<relation>` `<feature>`, where `<concept>` is a noun (e.g., "tiger"), `<Feature>` is a noun or adjective (e.g. "`stripe`") and `<relation>` is a verb representing the link between them (e.g. "have").

Some examples of valid triples are listed below.

```
tiger be animal
tiger live jungle
accordion produce music
accordion require air
```

Note that prepositions are not included in relations. So

```
tiger live jungle
accordion wear chest
airplane find airport
tiger use circus
```

would be true features, because tigers live in jungles, accordions are worn on chests, airplanes are found at airports, and tigers are used by circuses. You may assume absent prepositions when making your judgments.

Features need not be true of all instances of the concepts. So `tiger use circus` and `airplane do crash` are correct features, even though not all tigers are used by circuses and not all airplanes crash.

All feature terms have been made singular, so `tiger have tooth` and `accordion have key` would be correct triples, even though tigers have more than one tooth and accordions have more than one key.

Note that sometimes the concept noun is the agent of the relation verb, and sometimes the feature word is. So, for example, `airplane use passenger` is a valid feature of airplane, because passengers use airplanes.

Some features that you will see will be true, and some will be untrue. When judging the correctness of each `<concept> <relation> <feature>` triple, we would like you to select between four possibilities:

c: correct. Triple represents a correct, valid feature (e.g., `tiger be animal`, `airplane use passenger`)

p: plausible. Triple is not correct, but the triple may be plausible in a very specific set of circumstances (e.g., `tiger exhibit dimorphism`, `airplane land movie`) and/or the triple may be very general (e.g., `airplane be available`, `airplane have version`), or may be partly correct (e.g., `tiger be black`).

r: wrong, but related. The triple is wrong, but there is some kind of relationship between concept and the relation and/or feature (e.g., `motorcycle be car`, `accordion sing polka`, `accordion play astronaut`)

w: just completely wrong, for example `accordion fall Mississippi`, `tiger debunk reputation`.

Please use your own subjective judgment when making your decisions.

Appendix B

Human evaluation instructions: Part 2

In this part of the experiment, you are asked to judge whether semantic features—without relations—listed for concepts are true or not. The features are listed as word pairs of the form `<concept> -> <feature>`, where `<concept>` is a target concept noun (e.g., "`tiger`") and `<feature>` is a noun or adjective (e.g., "`stripe`"). Your task is to decide whether there is a relationship between the two words, and the strength of that relationship.

Some examples of valid pairs are listed below

```
tiger -> animal
tiger -> jungle
accordion -> music
accordion -> air
```

You may assume any correct/plausible relationship when making your judgments.

Note that features need not be true of all instances of the concepts. So `tiger -> circus and airplane ->` crash are correct features, even though not all tigers are used by circuses and not all airplanes crash.

Some pairs that you will see will represent true relationships, and some will be untrue. When judging the correctness of each `<concept> -> <feature>` pair, we would like you to select between four possibilities:

c: correct. Pair represents a correct, valid feature (e.g., `tiger -> animal, airplane -> passenger`)

p: plausible. Pair is not correct, but the pair may be plausible in a very specific set of circumstances (e.g., `tiger -> dimorphism, airplane -> terrorist`) and/or the pair may be very general (e.g., `airplane -> available, airplane -> large`), or may be partly correct (e.g., `tiger -> black`).

r: wrong, but related. The pair is wrong (i.e., not directly related), but there is some kind of tangential relationship between concept and the feature (e.g., `motorcycle -> screwdriver, airplane -> spaceship`).

w: just completely wrong, for example `airplane -> daisy, tiger -> lightbulb`.

Please use your own subjective judgment when making your decisions.