What is in a text, what isn't, and what this has to do with lexical semantics

Aurelie Herbelot
Universität Potsdam
aurelie.herbelot@cantab.net

Abstract

This paper queries which aspects of lexical semantics can reasonably be expected to be modelled by corpus-based theories such as distributional semantics or techniques such as ontology extraction. We argue that a full lexical semantics theory must take into account the *extensional* potential of words. We investigate to which extent corpora provide the necessary data to model this information and suggest that it may be partly learnable from text-based distributions, partly inferred from annotated data, using the insight that a concept's features are extensionally interdependent.

1 Introduction

Much work in computational linguistics relies on the use of corpora as evidential data, both for investigating language and for 'learning' about the world. Indeed, it is possible to inspect a great variety of phenomena, and get access to a lot of world knowledge, simply by having a large amount of text available. The purpose of this paper is to both acknowledge corpora as an invaluable data source for computational linguistics and point at their shortcomings. In particular, we want to argue that an appropriate representation of lexical meaning requires information beyond what is provided by written text and that this can be problematic for lexical models which rely entirely on corpora. Specifically, we wish to highlight how corpora fail to supply the information necessary to represent the *extension* of a term. In a nutshell, our argument is that the aspect of meaning represented by model theory is, in the best case, hard to extract and in the worst case not available at all when looking at corpus data. Building on this first discussion, we show that a concept's features are *extensionally* interdependent and, using this insight, propose that the part of model theory dealing with set relations (how much of set X is included in set Y?) may be learnt by exploiting a mixture of annotated (non-textual) data and standard distributional semantics. We present preliminary experiments investigating this hypothesis.

2 Text and extension

1. My cat – a mammal – likes sitting on the sofa.
2. There are more than two trees in this forest.
3. Cats have two eyes.

Sentences like 1–3 are not often found in corpora, or any kind of speech, for that matter. This is fortunate from the point of view of communication. In all three examples, the Gricean maxim of quantity is violated: 1 repeats information which is already implicitely given by the encyclopedic definition of the lexical item *cat*, while the other two express something obvious to anyone who has been in contact with forests/cats in their life (or indeed, who has read the encyclopedic definitions of *forest*/*cat*).

From the point of view of standard computational linguistics, this state of affairs is however undesirable. We will consider here two major approaches to extracting 'conceptual' knowledge from text:

distributional semantics and ontology extraction. Distributional semantics accounts for the lexical meaning of words by modelling, via appropriate weighting, their co-occurrence with other words (or any larger lexical context). The representation of a target word is thus a vector in a space where each dimension corresponds to a possible context. (Curran, 2003 and Turney and Pantel, 2010 are good overviews of this line of work). Ontology extraction, on the other hand, retrieves facts from text by searching for specific patterns: for instance, it is possible to find out that 'cat' is a hyponym of 'feline' by examining lexical patterns of the type 'X such as Y' (Hearst, 1992).

It should be clear that neither method can extract information which is simply absent from text. 3, for instance, may never occur in a corpus. This issue has been noted before, in particular in the psycholinguistics literature (Devereux et al., 2009; Andrews et al., 2009). To compensate, the use of manually obtained 'feature norms' is often advocated to complete corpus-based representations of concepts (e.g. a human might associate *bear* with the sensory-motor norms *claw, big, brown*, etc). But missing features are not the only issue. Let us return to 1. If we mentioned the word *mammal* everytime we speak of a cat, distributional semantics systems would be able to appropriately reflect the hyponymic relation between the two concepts – which in turn translates into an inclusion relation in model theory. But precisely because of the strong association between them, we (as good Griceans) are careful to leave the mammal out of the cat. Consequently, it is usual for a distribution to assign fairly low weights to features which we would argue are essential to the lexical representation of a word (see Geffet and Dagan, 2005 or Baroni et al., 2012 for attempts to capture the inclusion relation despite the flaws in the data).

Pattern-based methods obviously fare better when attempting to extract standard lexical relations. They do not, however, provide a true *extensional* analysis of the information they retrieve. For instance, from the snippet *four-legged animals such as cats and dogs*, we should not conclude that the set of dogs is included in the set of four-legged animals – a fair amount of dogs only have three legs. Our point is that relations obtained from corpora (or by coding feature norms) are essentially intensional: they do not model the part of semantics dealing with set relations and thus do not reflect our *expectations* with regard to a particular concept (i.e. how *likely* is it that a given bear is big, brown, etc?) We argue that such expectations are part of lexical semantics, as they mediate our use of words.[1] For instance, we strongly expect a dog to have four legs, but are not overly surprised when seeing a three-legged dog (the set of dogs is mostly, but not entirely, included in the set of four-legged things) and so would still call it a dog. Conversely, a forest with one tree is not intuitively a forest (the sets of trees making up each instance in the set of forests *all* have a 'large' cardinality).

In what follows we investigate how such set relations can be inferred from prior knowledge: we recall that features (whether of the distributional type or the sensory-motor type) are dependent on each other and that this dependency can be exploited to learn human-like expectations with regard to extension.

3 The extensional dependency hypothesis

Let us assume a particular kind of conceptual representation which consists of a vector of weighted features, as in distributional semantics, but where the weights are the *expected probabilities* of an instance of the concept to be associated with a particular distributional feature in the real world. So for instance, the feature *mammal* has a weight of 1 for the concept *cat* because all cats are mammals. In the same concept, *black* has perhaps a weight of 0.1 (assuming that one cat in ten is black). We call such representations **extensional distributions**, because each probability reflects some expectation about the inclusion relation between two sets.

Let us further assume a distributional space with n dimensions and let us refer to the extensional distribution of A as A°. We hypothesise that the value of A° along a dimension d_k is dependent on the value of A° along all other dimensions $d_{1\ldots n}$ in that space. Intuitively, this means that the probability that a cat (habitually) eats is dependent on the probability of that cat to (habitually) sleep, run, communicate, to be made of stone or to write books. In other words, the extensional distribution of a typical cat x

[1] Of course, expectations reflect certain common beliefs and no extensionally true facts about the world. But those beliefs can be modelled using the set-theory machinery, i.e. they can be regarded as a possible world.

reflects its status as a living (non-human) being, which in turn implies a high probability of cat$^\circ$ along the dimension *eat*. We call this the **extensional dependency hypothesis** (see Devereux et al., 2009 for related comments on this effect).

Such dependencies mean that learning the probabilities of certain features in relation to instances of a given target word can be greatly facilitated by learning another, information-rich feature. For instance, knowing that a is a bird allows us to infer many properties about a, e.g. that it lays eggs if it is female, that it probably flies, or that it perhaps builds nests.

In the absence of any supervision, it is hard to find out which features are most inference-producing. However, we can adopt a different strategy to learn inference rules. Let us assume that there *is* a correspondence between distributional data and the real world for at least some features. For example, we might find that as long as the predicate *be_v+fish_n* has been seen next to *pike* – be it only once –, we can infer that all pikes are fish. It is a property of the feature that if it applies to one individual of a kind, it will apply to all of them (contrast this with, e.g. *black_a*). If we can be reasonably certain that, given enough data, *be_v+fish_n* will be seen next to *pike*, we have a way to learn a real world probability from corpus data which we are confident in and which may be used for producing inferences. The rest of this paper investigates this hypothesis.

4 Experiments

4.1 A distributional system

Talking of '(expected) probabilities in the real world' has consequences in terms of choosing a particular notion of context for building our corpus-based distributions. Consider a system where context is defined by a word window around the target. *mouse* may be one of the terms sometimes appearing in the context of *cat*, but having a feature *mouse* does not tell us anything about how mice and cats are related in the real world (do mice eat cats? are made of cats? sell cats?) and so, the 'association' *mouse-cat* cannot be assigned a probability. In contrast, if we choose as context semantic dependencies of the type _+*eat*+*mouse* or _+*be*+*kept*+*as*+*pet*, where _ indicate the position of the target word, there is a clear interpretation of the context as being related to the target in the real world (what is the probability of a cat to eat mice? to be kept as a pet?) Consequently, we build a distributional system using dependency relations. Our data is the Wikiwoods corpus (Flickinger et al., 2010), a Wikipedia snapshot parsed with the English Resource Grammar (ERG), which we convert into a Dependency Minimal Recursion Semantics (DMRS, Copestake, 2009) format for our experiments. We only consider a number of dependency structures as context for a target word: adjectives, adverbs, prepositions and verbs, with their direct arguments, possessive constructs and coordination. The weight of a target word along a particular dimension of the semantic space is given by the normalised PMI measure proposed by Bouma (2007).

To speed up processing, we only retain dimensions which do occur with the lexical items in our experimental set (see 4.3 for a description). This results in our semantic space having 15799 dimensions.

4.2 The learning system

We design a system based on bootstrapping, which learns extensional distributions. The idea is to learn from our corpus the features which we are most confident in and use those to further the learning process.

Let us assume we wish to learn real-world probabilities for some features $F_1 ... F_n$, as applied to instances of various target words $w_k ... w_m$ (F_1 might be *lay_v+egg_n* while w_k might be *aardvark*). Let us also assume that we have some annotated data which provides us with 'rough' probabilities for a number of feature-instance pairs. For convenience of annotation – and because we do not expect humans to have an accurate probabilistic model of the world –, we express those probabilities using the natural language quantifiers *no, few, some, most, all* (see Herbelot and Copestake, 2011 for related work on resolving underspecified quantification). So we might already know that *most* swallows migrate.

The first iteration of our bootstrapping process runs over each feature F in $\{F_1 ... F_n\}$. A machine learning algorithm is fed the corpus-based distributions of the training instances (let us call them

$w_1...w_j$), together with the probabilities $p_m(F|w_1...j)$ from the annotated data. A classifier is learnt for F and its precision estimated by performing leave-one-out cross-validation on the training data. The classifier with the best precision, say $C(F_i)$ is recorded, together with its decisions, $p(F_i|w_1...j)$. The feature F_i is considered 'learnt'.

From the second iterations on, the following process takes place. For each feature F which has not yet been learnt, the machine learning algorithm is fed the corpus-based distributions of $w_1...w_j$, together with the values of the learnt features $p(F_{learnt}|w_1...j)$ and the manually annotated probabilities $p_m(F|w_1...j)$. As before, a classifier is learnt and its precision calculated by performing leave-one-out cross-validation on the training data and the classifier with the best precision, as well as its decisions on the training data, are recorded for use in the next iteration.

When classifying new, unseen instances, the classifiers are applied to the data in the order they were learnt during the training process. As an example, let us assume that we have learnt the probabilities of the features *be_v+fish_n* and *aquatic_a* for the animals *aardvark, cod* and *dolphin*. Let us also assume that *be_v+fish_n* was learnt first (it was classified with higher precision than *aquatic_a* in the first iteration of the algorithm). Let us further suppose that in the second iteration, the initial classifier for *aquatic_a* was modified to take into account the learnt feature *be_v+fish_n* (i.e. the precision of the classifier was increased by using the new information). Presented with a new classification problem, say *salmon*, the system would first try to find out the probability *p(be_v+fish_n|salmon)* and then use that figure to estimate *p(aquatic_a|salmon)*.

4.3 The dataset

We attempt to learn the quantification of a number of animal-feature pairs. The animals and features used in our experiment are chosen as follows.

The animals are picked by selecting the entries in the Wikipedia 'List of animals'[2] which have an occurrence count over 2000 in our corpus. This results in a set of 72 animals.

The features are chosen manually amongst the 50 most highly weighted vector components of ten different animal distributions. The animals considered are selected semi-randomly: we make sure that the most common types are included (mammals, fish, insects, birds, invertebrates). Features which satisfy the following conditions are included in the experiment: a) the feature must be applicable to the animal at the 'individual' level, i.e. it cannot be a temporary state of the animal, or apply to the species collectively ($black\_a$ is appropriate while $wounded\_a$ or $endangered\_a$ are not) b) the feature must be semantically 'complete', i.e. make sense when applied to the animal in isolation (*be_v+mammal_n* is appropriate while *behaviour_n+of_p* is not).

The feature selection exercise results in 54 vector components being selected.

Given our 72 animals and 54 features, we ask a human annotator to mark each animal-feature pair with a 'probability', expressed as a quantifier. Possible values are *no, few, some, most, all*. The guidelines for the annotation task can be seen at `http://www.cl.cam.ac.uk/~ah433/material/iwcs13-annot.pdf`.

5 Results

The following is based on an implementation of our system using Weka's[3] C4.5 (also referred to as J48) classifier (Quinlan, 1993). The C4.5 classifier produces a decision tree which can be inspected and is therefore particularly appropriate to 'check' learnt rules against human intuition. We perform leave-one-out validation on the first 10 animals in our dataset (for instance, we predict the values of our 54 features for *ant* using a training set consisting of all other animals). For each animal, the system therefore runs 1485 iterations of the classifier.

[2] `http://en.wikipedia.org/wiki/List_of_animals#Animals_.28by_common_name.29`

[3] `http://http://www.cs.waikato.ac.nz/~ml/weka/`

5.1 Finding relationships between corpora and the real world

In order to investigate whether real-world probabilities can be derived from corpus data, we first run a baseline system which classifies animal-feature pairs into our five quantifier classes, *no, few, some, most, all*, using only the corpus-based distributions of the animals in our training set.

The overall precision of the baseline system over our 540 predicted values is 0.51. As hypothesised, some features are learnt with high precision using word distributions only: *be_v+bird_n* is one such example, classifying all but one instances correctly (its estimated performance on the training data, as per cross-validation, is 0.95). Others, such as *carnivorous_a* consistently receive an incorrect classification (precision of 0 on the test data and 0.24 on the training data).

5.2 Learning inferences

We now turn to the question of learning rules which reflect real-world relationships. Because of space constraints, we are unable to reproduce actual output from the system, but examples can be seen at `http://www.cl.cam.ac.uk/~ah433/material/iwcs13-examples.pdf`. As illustration, we consider here the classifiers produced for the feature *aquatic_a*.

The baseline classifier (the one learnt from distributional data only) does use distributional features that are associated with water (*mediterranean_a*, *in_p()+water_n*), but overall, the decision tree is rather far from the way we might expect a human to attribute the feature *aquatic_a* to an animal species, and mostly includes seemingly irrelevant features such as *variant_of* or again *fascinating*.

The final classifier produced by our learning system, in contrast, involves the learnt probability for the feature *terrestrial_a*. It captures the facts that a) non-terrestrial species are aquatic and b) terrestrial species which live near water can also be said to be aquatic (cf. otters or beavers).

Other examples of real world dependencies learnt in the iterative process include *lay_v+egg_n* and *woolly_a* being dependent on *be_v+mammal_n*, *have_v+hair_n* depending on *woolly_a* and *be_v+insect_n*, or again *walk_v* depending on *be_v+mammal_n* and *fly_v*.

5.3 Overall performance

The overall precision of the iterative system is 0.48, which makes it lag behind the baseline by 3%. The result is disappointing but can be explained when inspecting the output. The strength of the system in catching meaningful dependencies between features, which we illustrated in the previous section, is also what makes it lose accuracy on our small dataset. Indeed, the dependencies integrated in the final classifiers assume that the learnt features were correctly predicted for the animal under consideration. This is unfortunately not the case, and errors accumulate during bootstrapping. For instance, when running on the test instance *ape*, the system classifies the features *woolly_a* and *lay_v+egg_n* by relying on the prediction for *be_v+mammal_n*. The precision of the system on that feature, however, is only 0.5. In order to truly evaluate the performance of the system, we suggest experimenting with a larger dataset.

6 Conclusion

This paper argued that there is a part of lexical semantics which is not dealt with by modelling techniques relying entirely on corpora. Indeed, standard speech tends to omit information which is true from an extensional point of view but irrelevant for successful communication. In order to retrieve the extensional potential of lexical terms, we proposed a separate, distribution-like representation which provides the *expected* real-world probabilities of instance-feature pairs and which we refer to as extensional distribution. We argued that a) for *some* features, there is a correspondence between corpus data and real world which can be learnt b) probabilities in the extensional distribution are dependent on each other and it is possible to infer unknown values from already learnt ones (including those learnt from corpus data).

Acknowledgement

The author is in receipt of a postdoctoral fellowship from the Alexander von Humboldt foundation.

References

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

Baroni, M., R. Bernardi, N.-Q. Do, and C.-c. Shan (2012). Entailment above the word level in distributional semantics. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL12)*.

Bouma, G. (2007). Normalized (Pointwise) Mutual Information in Collocation Extraction. *Proceedings of GSCL*, 31–40.

Copestake, A. (2009). Slacker semantics : why superficiality , dependency and avoidance of commitment can be the right way to go. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL09)*, Athens, Greece, pp. 1–9.

Curran, J. (2003). *From Distributional to Semantic Similarity*. Ph. D. thesis.

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

Flickinger, D., S. Oepen, and G. Ytrestol (2010). Wikiwoods: Syntacto-semantic annotation for english wikipedia. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC10)*.

Geffet, M. and I. Dagan (2005). The distributional inclusion hypothesises and lexical entailment. In *Proceedings Of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 107–114.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. Volume 14th International Conference on Computational Linguistics (COLING 92), pp. 539.

Herbelot, A. and A. Copestake (2011). Formalising and specifying underquantification. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, Oxford, United Kingdom, pp. 165–174.

Quinlan, J. R. (1993). *Programs for Machine Learning*. Morgan Kaufman.

Turney, P. D. and P. Pantel (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research 37*, 141–188.