Using Distributional Similarity to Identify Individual Verb Choice

Jing Lin
Department of Computing Science
University of Aberdeen
jlin@csd.abdn.ac.uk

Abstract

Human text is characterised by the individual lexical choices of a specific author. Significant variations exist between authors. In contrast, natural language generation systems normally produce uniform texts. In this paper we apply distributional similarity measures to help verb choice in a natural language generation system which tries to generate text similar to individual author. By using a distributional similarity (DS) measure on corpora collected from a recipe domain, we get the most likely verbs for individual authors. The accuracy of matching verb pairs produced by distributional similarity is higher than using the synonym outputs of verbs from WordNet. Furthermore, the combination of the two methods provides the best accuracy.

1 Introduction

Human text is characterised by the individual lexical choices of the specific author. It varies from author to author. Individual authors use different verbs to describe the same action. Natural language generation (NLG) systems, in contrast, normally produce uniform outputs without considering other lexical possibilities. Consider the following example from our corpora that are the BBC corpus and the Recipes for health eating corpus.

1. *BBC Corpus: Finely grate the ginger and squeeze out the juice into a shallow non-metallic dish. (BBC online recipes)*

2. *Author2: Extract juice from orange and add this with the water to the saucepan. (Recipes for health eating).*

Here, we can see that the two authors express the same type of action with different verbs, 'squeeze' and 'extract'. In fact, when expressing this action, the BBC corpus always use the verb 'squeeze', and Author2 only uses the verb 'extract'. Therefore, we can assume that Author2 considers the verb 'extract' to describe the same action as the verb 'squeeze' used by the BBC corpus. The purpose of our research is to develop a NLG system that can detect these kinds of individual writing features, such as the verb choice of individual authors, and can then generate personalised text.

The input of our personalised NLG system is an unseen recipe from the BBC food website. Our system, then, translates all sentences into the style of a personal author based on features drawn from analysing an individual corpus we collected. In this paper, we address the verb choice of the individual author in the translation process.

Our system defines the writing style of an individual author by analysing an individual corpus. Therefore, our system is a corpus-based NLG system. Lexical choice for individual authors is predicted by analysing the distributional similarity between words in a general large recipe corpus that is used to produce the verbs as the action representation and words in a specific indi-

Proceedings of the Fourth International Natural Language Generation Conference, pages 33–40,
Sydney, July 2006. ©2006 Association for Computational Linguistics

vidual recipe corpus. Firstly, we collected a large corpus in the recipe domain from the BBC online website. This large recipe corpus is used to extract feature values, for example verb choice, by analysing an individual corpus. Secondly, we collected our individual corpora for a number of individual authors. Each of them is used to extract feature values that may define the individual writing style. The individual author may choose the same or a different verb to describe cooking actions. The question is how can we identify the individual choice? For example, Author2 uses the verb 'extract' instead of the verb 'squeeze'. However, if the author does express the action by a different verb, the problem is how our system picks out verbs according to the individual choice of an author.

One way to solve this problem is to access large-scale manually constructed thesauri such as WordNet (Fellbaum, 1998), Roget's (Roget, 1911) or the Macquarie (Bernard, 1990) to get all synonyms and choose the most frequent one in the individual corpus. Another possible way is to use a lexical knowledge based system, like VerbNet (Kipper et al., 2000) to get more possible lexical choices. However, both methods only provide a number of pre-produced lexical choices that may or may not be the words that the individual author would choose. In other words, the lexical choice of an author may not be based on the synonyms extracted from one of the thesauri or may not even belong to the same semantic class. In our example, 'squeeze' and 'extract' are neither synonyms nor Coordinate Terms in WordNet. In a small domain, it is possible to manually build a verb list so that each action is described by a set of possible verbs. The drawback is that this is expensive. Furthermore, it still cannot catch verbs that are not included in the list. Is it possible to predict the individual verbs automatically?

The distributional hypothesis (Harris, 1968) says the following:

> The meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities.

Over recent years, many applications (Lin, 1998), (Lee, 1999), (Lee, 2001), (Weeds et al., 2004), and (Weeds and Weir, 2006) have been investigating the distributional similarity of words. Similarity means that words with similar meaning tend to appear in similar contexts. In NLG, the consideration of semantic similarity is usually preferred to just distributional similarity. However, in our case, the most important thing is to capture the most probable choice of a verb of an individual author for expressing an action. The expression of an action can be either the same verb, synonyms, or Coordinate terms to the verb in the big corpus, or any verbs that an individual author chooses for this action. If we check an individual corpus, there are a set of verbs in our list that do not occur. If these actions occur in the individual corpus, the individual author must use different verbs. Distributional similarity technology helps us to build the links between verbs in our list and the verbs in an individual corpus.

The rest of this paper is organised as follows. In Section 2, we describe the recipe domain, our corpora and our verb list. Section 3 disscuss our baseline system. In Section 4, we present the distributional similarity measures that we are proposing for analysing our corpora. The combination method is disscussed in Section5. In Section 6, we present an evaluation of our results. In Section 7, we draw conclusions and discuss future work.

2 The Recipe Domain and our Corpora

To find the most expressive verb pairs, we have to have corpora to be analysed. Therefore, the selection of a corpus is very important. As the research of authorship attribution (AA) shows (Burrow, 1987), (Holmes and Forsyth, 1995), (Keuelj et al., 2003), (Peng, 2003), and (Clement and Sharp, 2003), there can be style variations of an individual author. This happens even with the same topic and genre, and for the same action expressions. Firstly, a person's writing style can change as time, genre, and topic change. Can and Patton (Can and Patton, 2004) have drawn the conclusion:

> A higher time gap may have positive impact in separation and categorization.

Even within one text, the style may not be uniform. (Burrow, 1987) has pointed out that, for example, in fiction:

> The language of its dialogue and that of its narrative usually differ from each other in some obvious and many less obvious ways.

These problems require us to collect high-quality corpora. The recipe domain is a good start in this case. Sentences in it are narrative, imperative and objective, compared with other normal human text. For example, journal articles normally contain a large number of quotations, and they are more subjective. Furthermore, journal articles are more varied in content, even within the same topic. Secondly, most large corpora are not author-categorised. This requires us to collect our own individual corpora.

2.1 Our Corpora

As we mentioned before, we collected a general corpus in the recipe domain from the BBC food website. To make recipes varied enough, this corpus contains different cooking styles from western to eastern, different courses, including starters, main courses and desserts, and a number of recipes of famous cooks, such as Ainsley Harriott. Since recipes are widely-available both from the Internet and from publications, it is easy to collect author-categorised corpora. Our individual recipe corpora include four individual authors so far. Two of them are from two published recipe books, and another two we collected online. Recipe books are useful, because they are written in an unique style. Table 1 shows information about both our individual corpora and our large general corpus.

Our Corpora	**Number of Recipes**	**Total Lines**	**Total Words**
Large corpus (BBC online recipes)	823	6325	85594
Recipes for Health Eating	76	961	9212
Food for Health	113	1347	11791
CM (www.cooks.com)	48	537	6432
Jo Pratt (www.bbc.co.uk)	91	904	15417

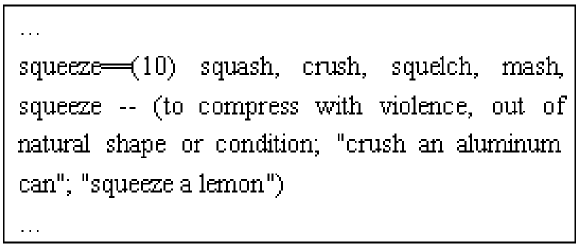
Table 1: Our corpora information

Although we are focusing on a small domain, verb variation between individual authors is a common phenomenon. Here are a few further examples from our corpora, which we want to capture.

1. *BBC corpus: Preheat the oven to 200C/400F/Gas 6. (BBC online food recipes)*
2. *Author2: Switch on oven to 200C, 400F or Gas Mark 6 and grease a $\frac{1}{2}$ litre ovenproof serving dish. (Recipes for Healthy Eating)*
3. *Author3. Put the oven on. (Food for Health)*

1. *BBC corpus: Sift the flour, baking powder and salt into a large bowl. (BBC online Recipes)*
2. *Author2: Sieve the flour, baking powder and bicarbonate of soda into a large mixing bowl. (Recipes for Health Eating)*
3. *Author3: Sieve the flour in, one-third at a time. (Food for Health)*

2.2 Our Verb List



...
squeeze==(10) squash, crush, squelch, mash, squeeze -- (to compress with violence, out of natural shape or condition; "crush an aluminum can"; "squeeze a lemon")
...

Figure 1: The information of the verblist

We manually built a verb list with 146 verbs in total from our BBC corpus. Each verb represents an unique cooking action, associated with definitions and synonyms extracted from WordNet. For example, the verb ’squeeze’ contains the following information shown in Figure 1. The BBC Corpus also contains a number of synonyms, such as the verb sift and the verb sieve. In this case, we only pick up the most frequent verb, which is the verb sift in this case, as an cooking action, and we record its synonyms, such as the verb sieve, in the late part of our verb list.

2.3 Using RASP in our corpora

Our data consists of verb-object pairs for verbs obtained from our BBC Corpus using RASP (Briscoe and Carroll, 2002). To derive reliable results, we deal with our data by the following rules. To avoid the sparse data problem and parsing mistakes, we removed a number of verbs that occur less than 3 times in our large corpus, and a set of mistake verbs made by the parser. We consider both direct objects and indirect objects together at the same time.

3 The Baseline Method - WordNet Synonyms

After the individual corpus is parsed, there are a number of main verbs appearing only in the BBC recipe corpus, but not in the individual corpus. This kind of main verbs is called *missing verb* in a corpus. For example, verbs such as 'roast', 'insert', 'drizzle' appear in the BBC corpus, but not in the Food for Health corpus. We say they are missing verbs in the Food for Health corpus. In this case, if the individual author expresses actions in the missing verb group, other verbs must be chosen instead. Our purpose is to find alternatives used by the individual author. To solve this problem, our baseline measure is the WordNet synonyms. If the missing verb contains synonyms in the verb list, we pick one as the candidate verb, called an *available candiate*. The following ways decide the verb alternatives for a missing verb. If there is more than one candidate verb for one missing verb, the most frequent synonym of the missing verb in the individual corpus is chosen as the alternative. The chosen synonym also has to be a main verb in the individual corpus. If the missing verb does not have a synonym or all available candidates do not appear in the individual corpus, we assign no alternative to this missing verb. In this case, we say there is no available alternative for the missing verb. The number of available alternatives for the missing verb and the accuracy is shown in Table 2, and Figure 2.

4 Distributional Similarity Measure

In this section, we introduce the idea of using distributional similarity measures, and discuss how this methodology can help us to capture verbs from individual authors.

By calculating the co-occurrence types of target words, distributional similarity defines the similarity between target word pairs. The *co-occurrence types* of a target word (w) are the context, c, in which it occurs and these have associated frequencies which may be used to form probability estimates (Weeds et al., 2004). In our case, the target word is main verbs of sentences and the co-occurrence types are objects. In section 6, similarity between verbs is derived from their objects, since normally there is no subject in the recipe domain. We are using the Additive t-test based Co-occurrence Retrieval Model of (Weeds and Weir, 2006). This method considers for each word *w* which co-occurrence types are retrieved. In our case, objects have been extracted from both the BBC Corpus and an individual corpus. Weeds and Weir use the the co-occurrence types as the features of word *(w)*, *F(w)*:

$$F(w) = \{c : D(w, c) > 0\}$$

where *D(w, c)* is the weight associated with word *w* and co-occurrence type *c*. T-test is used as a weight function, which is listed later.

Weeds and Weir use the following formula to describe the set of True Positives of co-occurrence types, which *w1* and *w2* are considered main verbs in copora:

$$TP(w_1, w_2) = F(w_1) \cap F(w_2)$$

They use the t-test from (Manning and Schütze, 1999) as the weight formula $D_t(w, c)$:

$$D_t(w, c) = \frac{p(c, w) - P(c)P(w)}{\sqrt{\frac{P(c,w)}{N}}}$$

Weeds and Weir then calculate the precision by using the proportion of features of *w1* which occurs in both words, and the recall by using the proportion of features of *w2* which occur in both words. In our experiment, precision is relative to

Individual Corpora	**Total Numbers of Missing Verbs**	**Available Candidates by WordNet**	**Available Candidates by DS**	**Available Candidates by Combination**	**Correct Alternatives by (DS VS. WordNet VS. Combination)**
Recipes for Health Eating	56	A = 36	A = 47	A = 52	8 VS. 10 VS. 17
Food for Health	57	A = 34	A = 52	A = 54	12 VS. 18 VS. 27
CM (www.cooks.com)	58	A = 25	A = 44	A = 51	10 VS. 4 VS. 14
Jo Pratt (www.bbc.co.uk)	26	A = 13	A = 22	A = 24	4 VS. 5 VS. 8

Table 2: The number of available missing verbs by the Distributional Similarity (DS) and by WordNet and by combination of DS and WordNet. ('A' means the total number of missing verbs in the individual corpus that have candidate alternatives in an individual corpus from methods.)

the BBC Corpus, and the recall is relative to an individual corpus.

$$P^{add}(w_1, w_2) = \frac{\sum_{TP(w_1,w_2)} D(w_1, c)}{\sum_{F(w_1)} D(w_1, c)}$$

$$R^{add}(w_1, w_2) = \frac{\sum_{TP(w_1,w_2)} D(w_2, c)}{\sum_{F(w_2)} D(w_{2,}c)}$$

Finally, Weeds and Weir combine precision and recall together by the following formulas:

$$m_h(P(w_1, w_2), R(w_1, w_2)) =$$

$$\frac{2.P(w_1, w_2).R(w_1, w_2)}{P(w_1, w_2) + R(w_1, w_2)}$$

$$m_a(P(w_1, w_2), R(w_1, w_2)) =$$

$$\beta.P(w_1, w_2) + (1 - \beta).R(w_1, w_2)$$

$$sim(w_1, w_2) = r.m_h(P(w_1, w_2), R(w_1, w_2))$$

$$+(1 - r).m_a(P(w_1, w_2), R(w_1, w_2))$$

where both *r*, β are between $[0, 1]$. In our experiments, we only assigned *r*=1. However, further performs can be done by assigning different values to *r* and β.

4.1 The Distributional Similarity method

Each missing verb in the BBC corpus is assigned the most likely verb as the available candidate from the individual corpus. The most likely verb is always chosen according to the largest similarity using the DS measure. In our case, if the largest similarity of the verb pair is larger than a certain value (-5.0), we say the missing verb has an *available candidate*. Otherwise, there is no available candidate existing in the individual corpus. For instance, DS suggests verb 'switch' is the most likely-exchangable verb for missing verb 'preheat' in the Recipes for Health Eating corpus. 'switch' appears 33 times in the individual corpus, in which there are 33 times that 'switch' has the same object as 'preheat'. Meanwhile, 'preheat' shows 191 times in total in the BBC corpus, with the same objects as 'switch' 176 times. By using the DS formulas, the similarity value between 'preheat' and 'switch' is 11.99. The number of available candidates of the missing verbs and the accuracy are shown in Table 2, and Figure 2.

There is only one corpus in the DS measures. In our case, *w1* and *w2* are from different corpora. For example, verb 'preheat' is from the BBC corpus, and verb 'switch' is in the Recipes for Health Eating. Although the co-occurence type is objects of the main verb, the precision is for the general corpus ——the BBC corpus, and the recall is for the individual corpus in our experiments.

5 The Combination method

We also combine the baseline and the DS method together in the combination method. The combination method tries the baseline first. For each missing verb, if the baseline returns an available alternative, this is the final available alternative of the combination method. If not, the available alternative is calculated by the DS method. If there is still no candidate for the missing verb, there is no available alternative in this case.

6 Evaluation

To justify accuracy of results by both the baseline method and the DS method, we manually judge whether or not the alternatives are inter changeable for the missing verbs. Table 2 shows the total number of missing verbs for each individual corpus and numbers of available alternatives as well. Also, it presents the number of correct alternatives for cases where both methods return answers, and results of a combination of two methods. In the future, we would like to evaluate the accuracy by more judges.

From Table 2, accuracies of distributional similarity are higher than WordNet synonyms in most cases, except in the individual corpus CM. The reason that CM got worse results is probably that the corpus size is not big enough. Since CM is the only individual corpus that has less than 50 recipes, this could lead to unreliable accuracy. In table 2, 'A' means the total number of missing verbs in the individual corpus that have candidate alternatives in an individual corpus from methods. It is obvious that distributional methods produce more available verbs than the synonyms of WordNet. In this case, we assume that WordNet is not very productive to provide alternative verb choices for individual authors compared with distributional similarity in a domain.

Figure 2 represents the accuracies of all methods. In Figure 2, we can see the overall accuracy of WordNet is not as good as the distributional similarity method. Moreover, we calculate the accuracy for the available verb pairs from the combination method of both the distributional similarity and WordNet. We can see that all combination accuracies are significantly better than accuracies of either distributional similarity or WordNet synonyms. In this case, distributional similarity and WordNet find different types of verbs. In other words, the similarity distributional method is very useful to find verbs that are not synonyms but represent the same type of action in individual corpora. And the type of verbs found by distributional similarity could not be pre-predicted, which makes the verb choice personalised.

In our verb pair outputs from distributional similarity, one problem is that we got similar verb pairs, for instances the verb 'simmer' matches to 'fry'. This is a common problem with distributional similarity, since it is not based on semantic meaning. This problem can perhaps be solved by building some hierarchical relationships between verbs. For instance, roast is one type of cooking.

The following examples are correct cases of verb pairs that are captured by distributional similarity. In each example, the semantic meanings of sentences are different, but the representation of action are the same.

roast (BBC Corpus) - cook (Food for Health):

1. *BBC Corpus: Season generously and roast for 30 minutes until softened and a little charred. (BBC online recipes)*

2. *Author2: Cover with a lid or foil and cook in the centre of the oven for 20 minutes, then turn down the oven to Reg 3 or 160C and continue cooking for 1 hour or until the kidneys are cooked. (Food for Health)*

saute (BBC Corpus) - fry (Food for Health):

1. *BBC Corpus: Melt the butter in a small to medium ovenproof pan and saute the cashew nuts for 2-3 minutes. (BBC online recipes)*

2. *Author2: Add the carrots and fry quickly for 5 minutes, stirring continuously. (Food for Health)*

preheat (BBC Corpus) - switch on (Food for Health):

1. *BBC Corpus: Preheat the oven to 200C/400F/Gas 6. (BBC online recipes)*

2. *Author2: Switch on oven to 190C, 375F or Gas Mark 5. (Food for Health)*

So far distributional similarity cannot capture the prepositions such as *on* in the third example. This is our future work.

7 Conclusion

In this paper, we used a distributional similarity method to help us to find matching verbs in

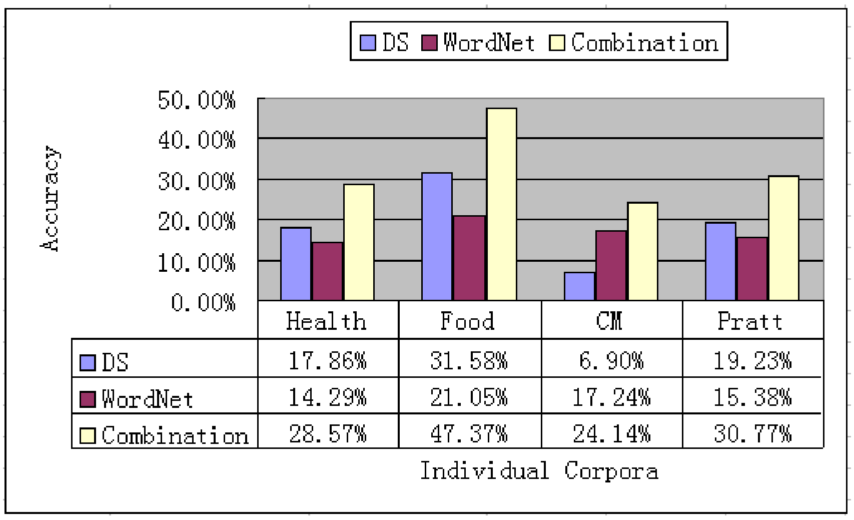


Figure 2: The Accuracy for Missing Verbs in Individual Corpora

an individual corpus. We have compared the result between the distributional similarity method and WordNet and the overall accuracy of distributional similarity is better than WordNet. Furthermore, the combination of the distributional similarity method and WordNet achieved the best accuracy. This suggests that distributional similarity is very helpful in choosing the proper verbs for individual authors. It is especially useful to find verbs that are not synonyms but represent the same type of action in individual corpora. This means distributional similarity can capture unpredicted verb preferences of individual authors from the individual corpora.

References

John R.L Bernard. 1990. *The Macquarie Encyclopedic Thesaurus*. The Macquarie Library, Sydney, Australia.

Edward Briscoe and John Carroll. 2002. Robust Accurate Statistical Annotation of General Text. In *Proceedings of the LREC-2002*, pages 1499–1504.

John F. Burrow. 1987. Word-patterns and Story-shapes: the Statistical Analusis of Narrative style. *Literary and Linguistic Computing*, 2(2):61–70.

Fazli Can and Jon M. Patton. 2004. Change of Writing Style with Time. *Computers and Humanities*, 38:61–82.

R. Clement and D. Sharp. 2003. Ngram and Bayesian Classification of Documents for Topic and Authorship. *Literary and Linguistic Computing*, 18(4):423–447.

Christiance Fellbaum, editor. 1998. *WordNet: An Electronic lexical Database*. MIT Press.

Zelig S. Harris. 1968. *Mathematical Structures of Language*. John Wiley.

D. I Holmes and R.S Forsyth. 1995. The Federalist Revisited: New Directions in Authoriship attribution. *Literary and Linguistic Computing*, 10(2):111–127.

V. Keuelj, F. C. Peng, N. Cercone, and C. Thomas. 2003. N-Gram-Based Author Profiles for Authorship Attribution. In *Proceedings the Pacific Association for Computational Linguistics.*

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-Based Construction of a Verb Lexicon. In *Proceedings the AAAI/IAAI*, pages 691–696.

Lillian Lee. 1999. Measure of Distributional Similarity. In *Proceedings of the Association for Computational Linguistics (ACL).*

Lillian Lee. 2001. On the Effectiveness of the Skew Divergence for Statistical Language. In *AIR*.

Dekang Lin. 1998. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the COLING-ACL.*

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

F. C. Peng. 2003. Language Independent Authorship Attribution using Character Level Language Models. In *Proceedings of the European Association for Computational Linguistics (ACL)*.

Peter Roget. 1911. *Thesaurus of English Words and Phrases*. Longmans.

Julie Weeds and David Weir. 2006. Co-occurrence Retrieval: A Flexible Framework for lexical Distributional Similarity. *Computational Linguistics*, 31(4):440–475.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Charactersing Measures of Lexical Distributional Similarity. In *Proceedings of the COLING*.