Learning the Countability of English Nouns from Corpus Data

Timothy Baldwin
CSLI
Stanford University
Stanford, CA, 12345
tbaldwin@csli.stanford.edu

Francis Bond
NTT Communication Science Laboratories
Nippon Telegraph and Telephone Corporation
Kyoto, Japan
bond@cslab.kecl.ntt.co.jp

Abstract

This paper describes a method for learning the countability preferences of English nouns from raw text corpora. The method maps the corpus-attested lexico-syntactic properties of each noun onto a feature vector, and uses a suite of memory-based classifiers to predict membership in 4 countability classes. We were able to assign countability to English nouns with a precision of 94.6%.

1 Introduction

This paper is concerned with the task of knowledge-rich lexical acquisition from unannotated corpora, focusing on the case of countability in English. Knowledge-rich lexical acquisition takes unstructured text and extracts out linguistically-precise categorisations of word and expression types. By combining this with a grammar, we can build broad-coverage deep-processing tools with a minimum of human effort. This research is close in spirit to the work of Light (1996) on classifying the semantics of derivational affixes, and Siegel and McKeown (2000) on learning verb aspect.

In English, nouns heading noun phrases are typically either **countable** or **uncountable** (also called **count** and **mass**). Countable nouns can be modified by denumerators, prototypically numbers, and have a morphologically marked plural form: *one dog*, *two dogs*. Uncountable nouns cannot be modified by denumerators, but can be modified by unspecific quantifiers such as *much*, and do not show any number distinction (prototypically being singular): **one equipment*, *some equipment*, **two equipments*. Many nouns can be used in countable or uncountable environments, with differences in interpretation.

We call the lexical property that determines which uses a noun can have the noun's countability preference. Knowledge of countability preferences is important both for the analysis and generation of English. In analysis, it helps to constrain the interpretations of parses. In generation, the countability preference determines whether a noun can become plural, and the range of possible determiners. Knowledge of countability is particularly important in machine translation, because the closest translation equivalent may have different countability from the source noun. Many languages, such as Chinese and Japanese, do not mark countability, which means that the choice of countability will be largely the responsibility of the generation component (Bond, 2001). In addition, knowledge of countability obtained from examples of use is an important resource for dictionary construction.

In this paper, we learn the countability preferences of English nouns from unannotated corpora. We first annotate them automatically, and then train classifiers using a set of gold standard data, taken from **COMLEX** (Grishman et al., 1998) and the transfer dictionaries used by the machine translation system **ALT-J/E** (Ikehara et al., 1991). The classifiers and their training are described in more detail in Baldwin and Bond (2003). These are then run over the corpus to extract nouns as members of four classes — countable: *dog*; uncountable: *furniture*; bipartite: *[pair of] scissors* and plural only: *clothes*.

We first discuss countability in more detail (§ 2). Then we present the lexical resources used in our experiment (§ 3). Next, we describe the learning process (§ 4). We then present our results and evaluation (§ 5). Finally, we discuss the theoretical and practical implications (§ 6).

2 Background

Grammatical countability is motivated by the semantic distinction between **object** and **substance** reference (also known as **bounded/non-bounded** or **individuated/non-individuated**). It is a subject of contention among linguists as to how far grammatical countability is semantically motivated and how much it is arbitrary (Wierzbicka, 1988).

In Proceedings of the 41st Annual Meeting of the Association for Computational
Linguistics, Sapporo, Japan, pp. 463-70.

The prevailing position in the natural language processing community is effectively to treat countability as though it were arbitrary and encode it as a lexical property of nouns. The study of countability is complicated by the fact that most nouns can have their countability changed: either converted by a lexical rule or embedded in another noun phrase. An example of conversion is the so-called universal packager, a rule which takes an uncountable noun with an interpretation as a substance, and returns a countable noun interpreted as a portion of the substance: *I would like two beers*. An example of embedding is the use of a classifier, e.g. uncountable nouns can be embedded in countable noun phrases as complements of classifiers: *one piece of equipment*.

Bond et al. (1994) suggested a division of countability into five major types, based on Allan (1980)'s noun countability preferences (NCPs). Nouns which rarely undergo conversion are marked as either fully countable, uncountable or plural only. Fully countable nouns have both singular and plural forms, and cannot be used with determiners such as *much, little, a little, less* and *overmuch*. Uncountable nouns, such as *furniture*, have no plural form, and can be used with *much*. Plural only nouns never head a singular noun phrase: *goods, scissors*.

Nouns that are readily converted are marked as either strongly countable (for countable nouns that can be converted to uncountable, such as *cake*) or weakly countable (for uncountable nouns that are readily convertible to countable, such as *beer*).

NLP systems must list countability for at least some nouns, because full knowledge of the referent of a noun phrase is not enough to predict countability. There is also a language-specific knowledge requirement. This can be shown most simply by comparing languages: different languages encode the countability of the same referent in different ways. There is nothing about the concept denoted by *lightning*, e.g., that rules out **a lightning* being interpreted as *a flash of lightning*. Indeed, the German and French translation equivalents are fully countable (*ein Blitz* and *un éclair* respectively). Even within the same language, the same referent can be encoded countably or uncountably: *clothes/clothing, things/stuff*, *jobs/work*.

Therefore, we must learn countability classes from usage examples in corpora. There are several impediments to this approach. The first is that words are frequently converted to different countabilities, sometimes in such a way that other native speakers will dispute the validity of the new usage. We do not necessarily wish to learn such rare examples, and may not need to learn more common conversions either, as they can be handled by regular lexical rules (Copestake and Briscoe, 1995). The second problem is that some constructions affect the apparent countability of their head: for example, nouns denoting a role, which are typically countable, can appear without an article in some constructions (e.g. *We elected him treasurer*). The third is that different senses of a word may have different countabilities: *interest* "a sense of concern with and curiosity" is normally countable, whereas *interest* "fixed charge for borrowing money" is uncountable.

There have been at several earlier approaches to the automatic determination of countability. Bond and Vatikiotis-Bateson (2002) determine a noun's countability preferences from its semantic class, and show that semantics predicts (5-way) countability 78% of the time with their ontology. O'Hara et al. (2003) get better results (89.5%) using the much larger Cyc ontology, although they only distinguish between countable and uncountable.

Schwartz (2002) created an automatic countability tagger (ACT) to learn noun countabilities from the British National Corpus. ACT looks at determiner co-occurrence in singular noun chunks, and classifies the noun if and only if it occurs with a determiner which can modify only countable or uncountable nouns. The method has a coverage of around 50%, and agrees with **COMLEX** for 68% of the nouns marked countable and with the **ALT-J/E** lexicon for 88%. Agreement was worse for uncountable nouns (6% and 44% respectively).

3 Resources

Information about noun countability was obtained from two sources. One was **COMLEX** 3.0 (Grishman et al., 1998), which has around 22,000 noun entries. Of these, 12,922 are marked as being countable (`COUNTABLE`) and 4,976 as being uncountable (`NCOLLECTIVE` or `:PLURAL *NONE*`). The remainder are unmarked for countability.

The other was the common noun part of **ALT-J/E**'s Japanese-to-English semantic transfer dictionary (Bond, 2001). It contains 71,833 linked Japanese-English pairs, each of which has a value for the noun countability preference of the English noun. Considering only unique English entries with different countability and ignoring all other information gave 56,245 entries. Nouns in the **ALT-J/E** dictionary are marked with one of the five major countability preference classes described in Section 2. In addition to countability, default values for number and classifier (e.g. *blade* for *grass*: *blade of grass*) are also part of the lexicon.

We classify words into four possible classes, with some words belonging to multiple classes. The first class is countable: **COMLEX**'s COUNTABLE and **ALT-J/E**'s fully, strongly and weakly countable. The second class is uncountable: **COMLEX**'s NCOLLECTIVE or :PLURAL *NONE* and **ALT-J/E**'s strongly and weakly countable and uncountable.

The third class is bipartite nouns. These can only be plural when they head a noun phrase (*trousers*), but singular when used as a modifier (*trouser leg*). When they are denumerated they use *pair*: *a pair of scissors*. **COMLEX** does not have a feature to mark bipartite nouns; *trouser*, for example, is listed as countable. Nouns in **ALT-J/E** marked plural only with a default classifier of *pair* are classified as bipartite.

The last class is plural only nouns: those that only have a plural form, such as *goods*. They can neither be denumerated nor modified by *much*. Many of these nouns, such as *clothes*, use the plural form even as modifiers (*a clothes horse*). The word *clothes* cannot be denumerated at all. Nouns marked :SINGULAR *NONE* in **COMLEX** and nouns in **ALT-J/E** marked plural only without the default classifier *pair* are classified as plural only. There was some noise in the **ALT-J/E** data, so this class was hand-checked, giving a total of 104 entries; 84 of these were attested in the training data.

Our classification of countability is a subset of **ALT-J/E**'s, in that we use only the three basic **ALT-J/E** classes of countable, uncountable and plural only, (although we treat bipartite as a separate class, not a subclass). As we derive our countability classifications from corpus evidence, it is possible to reconstruct countability preferences (i.e. fully, strongly, or weakly countable) from the relative token occurrence of the different countabilities for that noun.

In order to get an idea of the intrinsic difficulty of the countability learning task, we tested the **agreement** between the two resources in the form of classification accuracy. That is, we calculate the average proportion of (both positive and negative) countability classifications over which the two methods agree. E.g., **COMLEX** lists *tomato* as being only countable where **ALT-J/E** lists it as being both countable and uncountable. Agreement for this one noun, therefore, is 75.0%, as there is agreement for the classes of countable, plural only and bipartite (with implicit agreement as to negative membership for the latter two classes), but not for uncountable. Averaging over the total set of nouns countability-classified in both lexicons, the mean was 93.8%. Almost half of the disagreements came from words with two countabilities in **ALT-J/E** but only one in **COMLEX**.

4 Learning Countability

The basic methodology employed in this research is to identify lexical and/or constructional features associated with the countability classes, and determine the relative corpus occurrence of those features for each noun. We then feed the noun feature vectors into a classifier and make a judgement on the membership of the given noun in each countability class.

In order to extract the feature values from corpus data, we need the basic phrase structure, and particularly noun phrase structure, of the source text. We use three different sources for this phrase structure: part-of-speech tagged data, chunked data and fully-parsed data, as detailed below.

The corpus of choice throughout this paper is the written component of the British National Corpus (BNC version 2, Burnard (2000)), totalling around 90m w-units (POS-tagged items). We chose this because of its good coverage of different usages of English, and thus of different countabilities. The only component of the original annotation we make use of is the sentence tokenisation.

Below, we outline the features used in this research and methods of describing feature interaction, along with the pre-processing tools and extraction techniques, and the classifier architecture. The full range of different classifier architectures tested as part of this research, and the experiments to choose between them are described in Baldwin and Bond (2003).

4.1 Feature space

For each **target noun**, we compute a fixed-length feature vector based on a variety of features intended to capture linguistic constraints and/or preferences associated with particular countability classes. The feature space is partitioned up into **feature clusters**, each of which is conditioned on the occurrence of the target noun in a given construction.

Feature clusters take the form of one- or two-dimensional feature matrices, with each dimension describing a lexical or syntactic property of the construction in question. In the case of a one-dimensional feature cluster (e.g. noun occurring in singular or plural form), each component feature $feat_s$ in the cluster is translated into the 3-tuple:

$$\langle freq(feat_s|word), \frac{freq(feat_s|word)}{freq(word)}, \frac{freq(feat_s|word)}{\sum_i freq(feat_i|word)} \rangle$$

Feature cluster ***(base feature no.)***	***Countable***	***Uncountable***	***Bipartite***	***Plural only***
Head number (2)	S,P	S	P	P
Modifier number (2)	S,P	S	S	P
Subj–V agreement (2 × 2)	[S,S],[P,P]	[S,S]	[P,P]	[P,P]
Coordinate number (2 × 2)	[S,S],[P,S],[P,P]	[S,S],[S,P]	[P,S],[P,P]	[P,S],[P,P]
N *of* N (11 × 2)	[*100s*,P], …	[*lack*,S], …	[*pair*,P], …	[*rate*,P], …
PPs (52 × 2)	[*per*,-DET], …	[*in*,-DET], …	—	—
Pronoun (12 × 2)	[*it*,S],[*they*,P], …	[*it*,S], …	[*they*,P], …	[*they*,P], …
Singular determiners (10)	*a,each*, …	*much*, …	—	—
Plural determiners (12)	*many, few*, …	—	—	*many*, …
Neutral determiners (11 × 2)	[*less*,P], …	[BARE,S], …	[*enough*,P], …	[*all*,P], …

Table 1: Predicted feature-correlations for each feature cluster (S=singular, P=plural)

In the case of a two-dimensional feature cluster (e.g. subject-position noun number vs. verb number agreement), each component feature $feat_{s,t}$ is translated into the 5-tuple:

$$\langle freq(feat_{s,t}|word), \frac{freq(feat_{s,t}|word)}{freq(word)}, \frac{freq(feat_{s,t}|word)}{\sum_{i,j} freq(feat_{i,j}|word)}, \frac{freq(feat_{s,t}|word)}{\sum_{i} freq(feat_{i,t}|word)}, \frac{freq(feat_{s,t}|word)}{\sum_{j} freq(feat_{s,j}|word)}\rangle$$

See Baldwin and Bond (2003) for further details.

The following is a brief description of each feature cluster and its dimensionality (1D or 2D). A summary of the number of base features and prediction of positive feature correlations with countability classes is presented in Table 1.

Head noun number:^{1D} the number of the target noun when it heads an NP (e.g. *a shaggy dog* = SINGULAR)

Modifier noun number:^{1D} the number of the target noun when a modifier in an NP (e.g. *dog food* = SINGULAR)

Subject–verb agreement:^{2D} the number of the target noun in subject position vs. number agreement on the governing verb (e.g. *the dog barks* = ⟨SINGULAR,SINGULAR⟩)

Coordinate noun number:^{2D} the number of the target noun vs. the number of the head nouns of conjuncts (e.g. *dogs and mud* = ⟨PLURAL,SINGULAR⟩)

N *of* N constructions:^{2D} the number of the target noun (N_2) vs. the type of the N_1 in an N_1 *of* N_2 construction (e.g. *the type of dog* = ⟨TYPE,SINGULAR⟩). We have identified a total of 11 N_1 types for use in this feature cluster (e.g. COLLECTIVE, LACK, TEMPORAL).

Occurrence in PPs:^{2D} the presence or absence of a determiner (±DET) when the target noun occurs in **singular** form in a PP (e.g. *per dog* = ⟨*per*,-DET⟩). This feature cluster exploits the fact that countable nouns occur determinerless in singular form with only very particular prepositions (e.g. *by bus, *on bus, *with bus*) whereas with uncountable nouns, there are fewer restrictions on what prepositions a target noun can occur with (e.g. *on furniture*, *with furniture*, ?*by furniture*).

Pronoun co-occurrence:^{2D} what personal and possessive pronouns occur in the same sentence as singular and plural instances of the target noun (e.g. *The dog ate its dinner* = ⟨*its*,SINGULAR⟩). This is a proxy for pronoun binding effects, and is determined over a total of 12 third-person pronoun forms (normalised for case, e.g. *he, their, itself*).

Singular determiners:^{1D} what singular-selecting determiners occur in NPs headed by the target noun in singular form (e.g. *a dog* = *a*). All singular-selecting determiners considered are compatible with only countable (e.g. *another, each*) or uncountable nouns (e.g. *much, little*). Determiners compatible with either are excluded from the feature cluster (cf. *this dog, this information*). Note that the term "determiner" is used loosely here and below to denote an amalgam of simplex determiners (e.g. *a*), the null determiner, complex determiners (e.g. *all the*), numeric expressions (e.g. *one*), and adjectives (e.g. *numerous*), as relevant to the particular feature cluster.

Plural determiners:^{1D} what plural-selecting determiners occur in NPs headed by the target noun in plural form (e.g. *few dogs* = *few*). As with singular determiners, we focus on those plural-selecting determiners which are compatible with a proper subset of count, plural only and bipartite nouns.

Non-bounded determiners:^{2D} what non-bounded determiners occur in NPs headed by the target noun, and what is the number of the target noun for each (e.g. *more* *dogs* = ⟨*more*,PLURAL⟩). Here again, we restrict our focus to non-bounded determiners that select for singular-form uncountable nouns (e.g. *sufficient furniture*) and plural-form countable, plural only and bipartite nouns (e.g. *sufficient dogs*).

The above feature clusters produce a combined total of 1,284 individual feature values.

4.2 Feature extraction

In order to extract the features described above, we need some mechanism for detecting NP and PP boundaries, determining subject–verb agreement and deconstructing NPs in order to recover conjuncts and noun-modifier data. We adopt three approaches. First, we use part-of-speech (POS) tagged data and POS-based templates to extract out the necessary information. Second, we use chunk data to determine NP and PP boundaries, and medium-recall chunk adjacency templates to recover inter-phrasal dependency. Third, we fully parse the data and simply read off all necessary data from the dependency output.

With the POS extraction method, we first Penn-tagged the BNC using an fnTBL-based tagger (Ngai and Florian, 2001), training over the Brown and WSJ corpora with some spelling, number and hyphenation normalisation. We then lemmatised this data using a version of morph (Minnen et al., 2001) customised to the Penn POS tagset. Finally, we implemented a range of high-precision, low-recall POS-based templates to extract out the features from the processed data. For example, NPs are in many cases recoverable with the following Perl-style regular expression over Penn POS tags: `(PDT)* DT (RB|JJ[RS]?|NNS?)* NNS? [ˆN]`.

For the chunker, we ran fnTBL over the lemmatised tagged data, training over CoNLL 2000-style (Tjong Kim Sang and Buchholz, 2000) chunk-converted versions of the full Brown and WSJ corpora. For the NP-internal features (e.g. determiners, head number), we used the noun chunks directly, or applied POS-based templates locally within noun chunks. For inter-chunk features (e.g. subject–verb agreement), we looked at only adjacent chunk pairs so as to maintain a high level of precision.

As the full parser, we used RASP (Briscoe and Carroll, 2002), a robust tag sequence grammar-based parser. RASP's grammatical relation output function provides the phrase structure in the form of lemmatised dependency tuples, from which it is possible to read off the feature information. RASP has the advantage that recall is high, although precision is potentially lower than chunking or tagging as the parser is forced into resolving phrase attachment ambiguities and committing to a single phrase structure analysis.

Although all three systems map onto an identical feature space, the feature vectors generated for a given target noun diverge in content due to the different feature extraction methodologies. In addition, we only consider nouns that occur at least 10 times as head of an NP, causing slight disparities in the target noun type space for the three systems. There were sufficient instances found by all three systems for 20,530 common nouns (out of 33,050 for which at least one system found sufficient instances).

4.3 Classifier architecture

The classifier design employed in this research is four parallel supervised classifiers, one for each countability class. This allows us to classify a single noun into multiple countability classes, e.g. *demand* is both countable and uncountable. Thus, rather than classifying a given target noun according to the unique most plausible countability class, we attempt to capture its full range of countabilities. Note that the proposed classifier design is that which was found by Baldwin and Bond (2003) to be optimal for the task, out of a wide range of classifier architectures.

In order to discourage the classifiers from over-training on negative evidence, we constructed the gold-standard training data from unambiguously negative exemplars and potentially ambiguous positive exemplars. That is, we would like classifiers to judge a target noun as not belonging to a given countability class only in the absence of positive evidence for that class. This was achieved in the case of countable nouns, for instance, by extracting all countable nouns from each of the **ALT-J/E** and **COMLEX** lexicons. As positive training exemplars, we then took the intersection of those nouns listed as countable in both lexicons (irrespective of membership in alternate countability classes); negative training exemplars, on the other hand, were those contained in both lexicons but not classified as countable in either.[1] The uncountable gold-standard data

[1] Any nouns not annotated for countability in **COMLEX** were

Class	*Positive data*	*Negative data*	*Baseline*
Countable	4,342	1,476	.746
Uncountable	1,519	5,471	.783
Bipartite	35	5,639	.994
Plural only	84	5,639	.985

Table 2: Details of the gold-standard data

was constructed in a similar fashion. We used the **ALT-J/E** lexicon as our source of plural only and bipartite nouns, using all the instances listed as our positive exemplars. The set of negative exemplars was constructed in each case by taking the intersection of nouns not contained in the given countability class in **ALT-J/E**, with all annotated nouns with non-identical singular and plural forms in **COMLEX**.

Having extracted the positive and negative exemplar noun lists for each countability class, we filtered out all noun lemmata not occurring in the BNC.

The final make-up of the gold-standard data for each of the countability classes is listed in Table 2, along with a baseline classification accuracy for each class ("Baseline"), based on the relative frequency of the majority class (positive or negative). That is, for bipartite nouns, we achieve a 99.4% classification accuracy by arbitrarily classifying every training instance as negative.

The supervised classifiers were built using TiMBL version 4.2 (Daelemans et al., 2002), a memory-based classification system based on the k-nearest neighbour algorithm. As a result of extensive parameter optimisation, we settled on the default configuration for TiMBL with k set to 9. [2]

5 Results and Evaluation

Evaluation is broken down into two components. First, we determine the optimal classifier configuration for each countability class by way of stratified cross-validation over the gold-standard data. We then run each classifier in optimised configuration over the remaining target nouns for which we have feature vectors.

5.1 Cross-validated results

First, we ran the classifiers over the full feature set for the three feature extraction methods. In each case, we quantify the classifier performance by way of 10-fold stratified cross-validation over the gold-standard data for each countability class. The final classification accuracy and F-score[3] are averaged over the 10 iterations.

Class	*System*	*Accuracy (e.r.)*	*F-score*
Countable	Tagger*	.928 (.715)	.953
	Chunker	.933 (.734)	.956
	RASP*	.923 (.698)	.950
	Combined	.939 (.759)	.960
Uncountable	Tagger	.945 (.746)	.876
	Chunker*	.945 (.747)	.876
	RASP*	.944 (.743)	.872
	Combined	.952 (.779)	.892
Bipartite	**Tagger**	.997 (.489)	.752
	Chunker	.997 (.460)	.704
	RASP	.997 (.488)	.700
	Combined	.996 (.403)	.722
Plural only	Tagger	.989 (.275)	.558
	Chunker	.990 (.299)	.568
	RASP*	.989 (.227)	.415
	Combined	.990 (.323)	.582

Table 3: Cross-validation results

The cross-validated results for each classifier are presented in Table 3, broken down into the different feature extraction methods. For each, in addition to the F-score and classification accuracy, we present the relative error reduction (e.r.) in classification accuracy over the majority-class baseline for that gold-standard set (see Table 2). For each countability class, we additionally ran the classifier over the concatenated feature vectors for the three basic feature extraction methods, producing a 3,852-value feature space ("Combined").

Given the high baseline classification accuracies for each gold-standard dataset, the most revealing statistics in Table 3 are the error reduction and F-score values. In all cases other than bipartite, the combined system outperformed the individual systems. The difference in F-score is statistically significant (based on the two-tailed t-test, $p < .05$) for the asterisked systems in Table 3. For the bipartite class, the difference in F-score is not statistically significant between any system pairing.

There is surprisingly little separating the tagger-, chunker- and RASP-based feature extraction methods. This is largely due to the precision/recall trade-off noted above for the different systems.

5.2 Open data results

We next turn to the task of classifying all unseen common nouns using the gold-standard data and the best-performing classifier configurations for each countability class (indicated in bold in Table 3).[4]

ignored in this process so as to assure genuinely negative exemplars.

[2] We additionally experimented with the kernel-based TinySVM system, but found TiMBL to be superior in all cases.

[3] Calculated according to: $\frac{2 \cdot precision \cdot recall}{precision + recall}$

[4] In each case, the classifier is run over the best-500 features as selected by the method described in

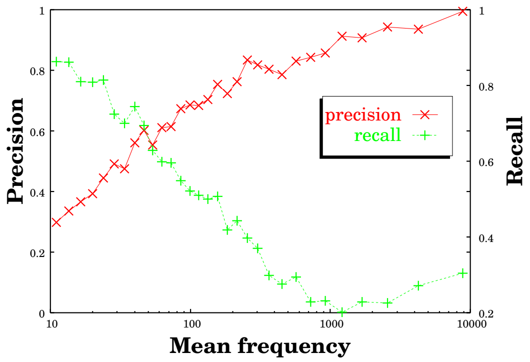

Figure 1: Precision–recall curve for countable nouns

Here, the baseline method is to classify every noun as being uniquely countable.

There were 11,499 feature-mapped common nouns not contained in the union of the gold-standard datasets. Of these, the classifiers were able to classify 10,355 (90.0%): 7,974 (77.0%) as countable (e.g. *alchemist*), 2,588 (25.0%) as uncountable (e.g. *ingenuity*), 9 (0.1%) as bipartite (e.g. *headphones*), and 80 (0.8%) as plural only (e.g. *damages*). Only 139 nouns were assigned to multiple countability classes.

We evaluated the classifier outputs in two ways. In the first, we compared the classifier output to the combined **COMLEX** and **ALT-J/E** lexicons: a lexicon with countability information for 63,581 nouns. The classifiers found a match for 4,982 of the nouns. The predicted countability was judged correct 94.6% of the time. This is marginally above the level of match between **ALT-J/E** and **COMLEX** (93.8%) and substantially above the baseline of all-countable at 89.7% (error reduction = 47.6%).

To gain a better understanding of the classifier performance, we analysed the correlation between corpus frequency of a given target noun and its precision/recall for the countable class.[5] To do this, we listed the 11,499 unannotated nouns in increasing order of corpus occurrence, and worked through the ranking calculating the mean precision and recall over each partition of 500 nouns. This resulted in the precision–recall graph given in Figure 1, from which it is evident that mean recall is proportional and precision inversely proportional to corpus frequency. That is, for lower-frequency nouns, the classifier tends to rampantly classify nouns as countable, while for higher-frequency nouns, the classifier tends to be extremely conservative in positively classifying nouns. One possible explanation for this is that, based on the training data, the frequency of a noun is proportional to the number of countability classes it belongs to. Thus, for the more frequent nouns, evidence for alternate countability classes can cloud the judgement of a given classifier.

In secondary evaluation, the authors used BNC corpus evidence to blind-annotate 100 randomly-selected nouns from the test data, and tested the correlation with the system output. This is intended to test the ability of the system to capture corpus-attested usages of nouns, rather than independent lexicographic intuitions as are described in the **COMLEX** and **ALT-J/E** lexicons. Of the 100, 28 were classified by the annotators into two or more groups (mainly countable and uncountable). On this set, the baseline of all-countable was 87.8%, and the classifiers gave an agreement of 92.4% (37.7% e.r.), agreement with the dictionaries was also 92.4%. Again, the main source of errors was the classifier only returning a single countability for each noun. To put this figure in proper perspective, we also hand-annotated 100 randomly-selected nouns from the training data (that is words in our combined lexicon) according to BNC corpus evidence. Here, we tested the correlation between the manual judgements and the combined **ALT-J/E** and **COMLEX** dictionaries. For this dataset, the baseline of all-countable was 80.5%, and agreement with the dictionaries was a modest 86.8% (32.3% e.r.). Based on this limited evaluation, therefore, our automated method is able to capture corpus-attested countabilities with greater precision than a manually-generated static repository of countability data.

6 Discussion

The above results demonstrate the utility of the proposed method in learning noun countability from corpus data. In the final system configuration, the system accuracy was 94.6%, comparing favourably with the 78% accuracy reported by Bond and Vatikiotis-Bateson (2002), 89.5% of O'Hara et al. (2003), and also the noun token-based results of Schwartz (2002).

At the moment we are merely classifying nouns into the four classes. The next step is to store the distribution of countability for each target noun and build a representation of each noun's countability preferences. We have made initial steps in this direction, by isolating token instances strongly support-

Baldwin and Bond (2003) rather than the full feature set, purely in the interests of reducing processing time. Based on cross-validated results over the training data, the resultant difference in performance is not statistically significant.

[5] We similarly analysed the uncountable class and found the same basic trend.

ing a given countability class analysis for that target noun. We plan to estimate the overall frequency of the different countabilities based on this evidence. This would represent a continuous equivalent of the discrete 5-way scale employed in **ALT-J/E**, tunable to different corpora/domains.

For future work we intend to: investigate further the relation between meaning and countability, and the possibility of using countability information to prune the search space in word sense disambiguation; describe and extract countability-idiosyncratic constructions, such as determinerless PPs and role-nouns; investigate the use of a grammar that distinguishes between countable and uncountable uses of nouns; and in combination with such a grammar, investigate the effect of lexical rules on countability.

7 Conclusion

We have proposed a knowledge-rich lexical acquisition technique for multi-classifying a given noun according to four countability classes. The technique operates over a range of feature clusters drawing on pre-processed corpus data, which are then fed into independent classifiers for each of the countability classes. The classifiers were able to selectively classify the countability preference of English nouns with a precision of 94.6%.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. BCS-0094638 and also the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University. We would like to thank Leonoor van der Beek, Ann Copestake, Ivan Sag and the three anonymous reviewers for their valuable input on this research.

References

Keith Allan. 1980. Nouns and countability. *Language*, 56(3):541–67.

Timothy Baldwin and Francis Bond. 2003. A plethora of methods for learning English countability. In *Proc. of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, Sapporo, Japan. (to appear).

Francis Bond and Caitlin Vatikiotis-Bateson. 2002. Using an ontology to determine English countability. In *Proc. of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan.

Francis Bond, Kentaro Ogura, and Satoru Ikehara. 1994. Countability and number in Japanese-to-English machine translation. In *Proc. of the 15th International Conference on Computational Linguistics (COLING '94)*, pages 32–8, Kyoto, Japan.

Francis Bond. 2001. *Determiners and Number in English, contrasted with Japanese, as exemplified in Machine Translation*. Ph.D. thesis, University of Queensland, Brisbane, Australia.

Ted Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1499–1504, Las Palmas, Canary Islands.

Lou Burnard. 2000. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

Ann Copestake and Ted Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of Semantics*, pages 15–67.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2002. *TiMBL: Tilburg memory based learner, version 4.2, reference guide*. ILK technical report 02-01.

Ralph Grishman, Catherine Macleod, and Adam Myers, 1998. *COMLEX Syntax Reference Manual*. Proteus Project, NYU. (`http://nlp.cs.nyu.edu/comlex/refman.ps`).

Satoru Ikehara, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. 1991. Toward an MT system without pre-editing – effects of new methods in **ALT-J/E**–. In *Proc. of the Third Machine Translation Summit (MT Summit III)*, pages 101–106, Washington DC.

Marc Light. 1996. Morphological cues for lexical semantics. In *Proc. of the 34th Annual Meeting of the ACL*, pages 25–31, Santa Cruz, USA.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–23.

Grace Ngai and Radu Florian. 2001. Transformation-based learning in the fast lane. In *Proc. of the 2nd Annual Meeting of the North American Chapter of Association for Computational Linguistics (NAACL2001)*, pages 40–7, Pittsburgh, USA.

Tom O'Hara, Nancy Salay, Michael Witbrock, Dave Schneider, Bjoern Aldag, Stefano Bertolo, Kathy Panton, Fritz Lehmann, Matt Smith, David Baxter, Jon Curtis, and Peter Wagner. 2003. Inducing criteria for mass noun lexical mappings using the Cyc KB and its extension to WordNet. In *Proc. of the Fifth International Workshop on Computational Semantics (IWCS-5)*, Tilburg, the Netherlands.

Lane O.B. Schwartz. 2002. *Corpus-based acquisition of head noun countability features*. Master's thesis, Cambridge University, Cambridge, UK.

Eric V. Siegel and Kathleen McKeown. 2000. Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4):595–627.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task: Chunking. In *Proc. of the 4th Conference on Computational Natural Language Learning (CoNLL-2000)*, Lisbon, Portugal.

Anna Wierzbicka. 1988. *The Semantics of Grammar*. John Benjamin.