Identifying non-referential *it*: a machine learning approach incorporating linguistically motivated patterns

Adriane Boyd

Department of Linguistics
The Ohio State University
1712 Neil Ave.
Columbus, OH 43210
adriane@ling.osu.edu

Whitney Gegg-Harrison & Donna Byron

Department of Computer Science and Engineering
The Ohio State University
2015 Neil Ave.
Columbus, OH 43210
{geggharr,dbyron}@cse.osu.edu

Abstract

In this paper, we present a machine learning system for identifying non-referential *it*. Types of non-referential *it* are examined to determine relevant linguistic patterns. The patterns are incorporated as features in a machine learning system which performs a binary classification of *it* as referential or non-referential in a POS-tagged corpus. The selection of relevant, generalized patterns leads to a significant improvement in performance.

1 Introduction

The automatic classification of *it* as either referential or non-referential is a topic that has been relatively ignored in the computational linguistics literature, with only a handful of papers mentioning approaches to the problem. With the term "non-referential *it*", we mean to refer to those instances of *it* which do not introduce a new referent. In the previous literature these have been called "pleonastic", "expletive", and "non-anaphoric". It is important to be able to identify instances of non-referential *it* to generate the correct semantic interpretation of an utterance. For example, one step of this task is to associate pronouns with their referents. In an automated pronoun resolution system, it is useful to be able to skip over these instances of *it* rather than attempt an unnecessary search for a referent for them,

only to end up with inaccurate results. The task of identifying non-referential *it* could be incorporated into a part-of-speech tagger or parser, or viewed as an initial step in semantic interpretation.

We develop a linguistically-motivated classification for non-referential *it* which includes four types of non-referential *it*: extrapositional, cleft, weather/condition/time/place, and idiomatic, each of which will be discussed in more detail in Section 2. A subset of the BNC Sampler Corpus (Burnard, 1995) was chosen for our task because of its extended tagset and high tagging accuracy. Non-referential *it* makes up a significant proportion of the occurrences of *it* in our corpus, which contains a selection of written texts of various genres, approximately one-third prose fiction, one-third newspaper text, and one-third other non-fiction. In our corpus, there are 2337 instances of *it*, 646 of which are non-referential (28%). *It* appears in over 10% of the sentences in our corpus. The corpus is described in further detail in Section 3.

Previous research on this topic is fairly limited. Paice and Husk (1987) introduces a rule-based method for identifying non-referential *it* and Lappin and Leass (1994) and Denber (1998) describe rule-based components of their pronoun resolution systems which identify non-referential *it*. Evans (2001) describes a machine learning system which classifies *it* into seven types based on the type of referent. Their approaches are described in detail in Section 4. In Section 5 we describe our system which combines and extends elements of the systems developed by Paice and Husk (1987) and Evans (2001), and the results are presented in Section 6.

2 Classification

The first step is to create a classification system for all instances of *it*. Though the goal is the binary classification of *it* as referential or non-referential, an annotation scheme is used which gives more detail about each instance of non-referential *it*, since they occur in a number of constructions. The main types of non-referential *it* are taken from the *Cambridge Grammar of the English Language* in the section on "Special uses of *it*", Section 2.5, Huddleston and Pullum (2002). Five main uses are outlined: extrapositional, cleft, weather/condition/time/place, idiomatic, and predicative. As noted in the *Cambridge Grammar*, predicative *it* seems to be more referential that the other types of non-referential *it*. Predicative *it* can typically be replaced with a demonstrative pronoun. Consider the example: *It is a dreary day. It* can be replaced with *This* with no change in grammaticality and no significant change in meaning: *This is a dreary day.* In contrast, replacing the other types of *it* with *this* results in nonsense, e.g., **This seems that the king is displeased.*

For our purposes, if a particular *it* can be replaced with a demonstrative pronoun and the resulting sentence is still grammatical and has no significant change in meaning, this *it* is referential and therefore annotated as referential. The demonstrative pronoun replacement test is not quite perfect (e.g., **This is a dreary day in Paris*), but no such instances of predicative *it* were found in the corpus so predicative *it* is always classified as referential. This leaves four types of *it*, each of which are described in detail below. The main examples for each type are taken from the corpus. See Section 3 for details about the corpus.

2.1 Extrapositional

When an element of a sentence is extraposed, *it* is often inserted as a placeholder in the original position of the now extraposed element. Most often, *it* appears in the subject position, but it can also appear as an object. Example (1) lists a few instances of extrapositional *it* from our corpus.

(1) a. **It** has been confirmed this week that political parties will no longer get financial subsidies.

b. She also made **it** clear that Conductive Education is not the only method.

c. You lead life, **it** seems to me, like some ritual that demands unerring performance.

The extraposed element is typically a subordinate clause, and the type of clause depends on lexical properties of verbs and adjectives in the sentence, see (2).

(2) * It was difficult that X.
    It was difficult to X.
   * It was clear to X.
    It was clear that X.

As (1c) shows, extrapositional *it* can also appear as part of a truncated extrapositional phrase as a kind of parenthetical comment embedded in a sentence.

2.2 Cleft

It appears as the subject of *it*-cleft sentences. When an *it*-cleft sentence is formed, the foregrounded phrase becomes the complement of the verb *be* and the rest of sentence is backgrounded in a relative clause. The foregrounded phrase in a cleft sentence can be a noun phrase, prepositional phrase, adjective phrase, adverb phrase, non-finite clause, or content clause.

(3) a. It was **the military district commander** who stepped in to avoid bloodshed. (*noun phrase*)

b. It is **on this point** that the views of the SACP and some Soviet policymakers divide. (*prepositional phrase*)

c. 'Tis **glad** I am to 'ear it, me lord. (*adjective phrase*)

Additionally, the foregrounded phrase can sometimes be fronted:

(4) **He** it was who ushered in the new head of state.

More context than the immediate sentence is needed to accurately identify *it*-cleft sentences. First, clefts with a foregrounded noun phrase are ambiguous between cleft sentences (5a) and sentences where the noun phrase and relative clause form a constituent (5b).

(5) a. A: I heard that the general stepped in to avoid bloodshed.
      B: No, it was the military district commander who stepped in.

    b. A: Was that the general being interviewed on the news?
      B: No, it was the military district commander who stepped in to avoid bloodshed.

Due to this ambiguity, we expect that it may be difficult to classify clefts. In addition, there are difficulties because the relative clause does not always appear in full. In various situations the relative pronoun can be omitted, the relative clause can be reduced, or the relative clause can be omitted entirely.

2.3 Weather/Condition/Time/Place

It appears as the subject of weather and other related predicates involving condition, time, and place/distance:

(6) a. It was snowing steadily outside.

    b. It was about midnight.

    c. It was no distance to Mutton House.

    d. It was definitely not dark.

2.4 Idiomatic

In idioms, *it* can appear as the subject, object, or object of a preposition.

(7) a. After three weeks it was my turn to go to the delivery ward at Fulmer.

    b. Cool it!

    c. They have not had an easy time of it.

2.5 General Notes

Non-referential *it* is most often the subject of a sentence, but in extrapositional and idiomatic cases, it can also be the object. Idioms are the only cases where non-referential *it* is found as the object of a preposition.

3 Corpus

The BNC Sampler Corpus (Burnard, 1995) was chosen for its extended tagset and high tagging accuracy. The C7 tagset used for this corpus has a unique

Prose fiction	32%
Newspaper text	38%
Other non-fiction	30%

Table 1: Text types in our corpus

	# of Instances	% of Inst.
Extrapositional	477	20.4%
Cleft	119	5.1%
Weather	69	2.9%
Idiomatic	46	2.0%
Referential	1626	69.6%
Total	2337	100%

Table 2: Instances of *it* in our corpus

tag for *it*, which made the task of identifying all occurrences of *it* very simple. We chose a subset consisting of 350,000 tokens from written texts in variety of genres. The breakdown by text type can be seen in Table 1.

The two lead authors independently annotated each occurence with one of the labels shown in Table 2 and then came to a joint decision on the final annotation. The breakdown of the instances of *it* in our corpus is shown in Table 2. There are 2337 occurrences of *it*, 646 of which are non-referential (28%). Ten percent of the corpus, taken from all sections, was set aside as test data. The remaining section, which contains 2100 instances of *it*, became our training data.

4 Previous Research

Paice and Husk (1987) reports a simple rule-based system that was used to identify non-referential *it* in the technical section of the Lancaster-Oslo/Bergen Corpus. Because of the type of text, the distribution of types of non-referential *it* is somewhat limited, so they only found it necessary to write rules to match extrapositional and cleft *it* (although they do mention two idioms found in the corpus). The corpus was plain text, so their rules match words and punctuation directly.

Their patterns find *it* as a left bracket and search for a right bracket related to the extrapositional and cleft grammatical patterns (*to*, *that*, etc.). For the extrapositional instances, there are lists of words which are matched in between *it* and the right

Accuracy	92%
Precision	93%
Recall	97%

Table 3: Paice and Husk (1987): Results

Accuracy	79%
Precision	80%
Recall	31%

Table 4: Replicating Paice and Husk (1987)

Accuracy	71%
Precision	73%
Recall	69%

Table 5: Evans (2001): Results, Binary Classification

bracket. The word lists are task-status words (STATUS), state-of-knowledge words (STATE), and a list of prepositions and related words (PREP), which is used to rule out right brackets that could potentially be objects of prepositions. Patterns such as "it STATUS to" and "it !PREP that" were created. The left bracket can be at most 27 words from the right bracket and there can be either zero or two or more commas or dashes between the left and right brackets. Additionally, their system had a rule to match parenthetical *it*: there is a match when *it* appears immediately following a comma and another comma follows within four words. Their results, shown in Table 3, are impressive.

We replicated their system and ran it on our testing data, see Table 4. Given the differences in text types, it is not surprising that their system did not perform as well on our corpus. The low recall seems to show the limitations of fixed word lists, while the reasonably high precision shows that the simple patterns tend to be accurate in the cases where they apply.

Lappin and Leass (1994) and Denber (1998) mention integrating small sets of rules to match non-referential *it* into their larger pronoun resolution systems. Lappin and Leass use two words lists and a short set of rules. One word list is modal adjectives (*necessary*, *possible*, *likely*, etc.) and the other is cognitive verbs (*recommend*, *think*, *believe*, etc.). Their rules are as follows:

It is Modaladj that S
It is Modaladj (for NP) to VP
It is Cogv-ed that S
It seems/appears/means/follows (that) S
NP makes/finds it Modaladj (for NP) to VP

It is time to VP
It is thanks to NP that S

Their rules are mainly concerned with extrapositional *it* and they give no mention of cleft *it*. They give no direct results for this component of their system, so it is not possible to give a comparison. Denber (1998) includes a slightly revised and extended version of Lappin and Leass's system and adds in detection of weather/time *it*. He suggests using WordNet to extend word lists.

Evans (2001) begins by noting that a significant percentage of instances of *it* do not have simple nominal referents and describes a system which uses a memory-based learning (MBL) algorithm to perform a 7-way classification of *it* by type of referent. We consider two of his categories, *pleonastic* and *stereotypic/idiomatic*, to be non-referential. Evans created a corpus with texts from the BNC and SUSANNE corpora and chose to use a memory-based learning algorithm. A memory-based learning algorithm classifies new instances on the basis of their similarity to instances seen in the training data. Evans chose the k-nearest neighbor algorithm from the Tilburg Memory-Based Learner (TiMBL) package (Daelemans et al., 2003) with approximately 35 features relevant to the 7-way classification. Although his system was created for the 7-way classification task, he recognizes the importance of the binary referential/non-referential distinction and gives the results for the binary classification of pleonastic *it*, see Table 5. His results for the classification of idiomatic *it* (33% precision and 0.7% recall) show the limitations of a machine learning system given sparse data.

We replicated Evans's system with a simplified set of features to perform the referential/non-referential classification of *it*. We did not include features that would require chunking or features that seemed relevant only for distinguishing kinds of referential *it*. The following thirteen features are used:

Accuracy	76%
Precision	57%
Recall	60%

Table 6: Replicating Evans (2001)

1-8. four preceding and following POS tags

9-10. lemmas of the preceding and following verbs

11. lemma of the following adjective

12. presence of *that* following

13. presence of an immediately preceding preposition

Using our training and testing data with the same algorithm from TiMBL, we obtained results similar to Evans's, shown in Table 6. The slightly higher accuracy is likely due to corpus differences or the reduced feature set which ignores features largely relevant to other types of *it*.

Current state-of-the-art reference resolution systems typically include filters for non-referential noun phrases. An example of such a system is Ng and Cardie (2002), which shows the improvement in reference resolution when non-referential noun phrases are identified. Results are not given for the specific task of identifying non-referential *it*, so a direct comparison is not possible.

5 Method

As seen in the previous section, both rule-based and machine learning methods have been shown to be fairly effective at identifying non-referential *it*. Rule-based methods look for the grammatical patterns known to be associated with non-referential *it* but are limited by fixed word lists; machine learning methods can handle open classes of words, but are less able to generalize about the grammatical patterns associated with non-referential *it* from a small training set.

Evans's memory-based learning system showed a slight integration of rules into the machine learning system by using features such as the presence of following *that*. Given the descriptions of types of non-referential *it* from Section 2, it is possible to create more specific rules which detect the fixed grammatical patterns associated with non-referential *it* such as *it VERB that* or *it VERB ADJ to*. Many of these

patterns are similar to Paice and Husk's, but having part-of-speech tags allows us to create more general rules without reference to specific lexical items. If the results of these rule matches are integrated as features in the training data for a memory-based learning system along with relevant verb and adjective lemmas, it becomes possible to incorporate knowledge about grammatical patterns without creating fixed word lists. The following sections examine each type of non-referential *it* and describe the patterns and features that can be used to help automatically identify each type.

5.1 Extrapositional *it*

Extrapositional *it* appears in a number of fairly fixed patterns, nine of which are shown below. Intervening tokens are allowed between the words in the patterns. **F4-6** are more general versions of **F1-3** but are not as indicative of non-referential *it*, so it useful to keep them separate even though ones that match **F1-3** will also match **F4-6**. **F7** applies when *it* is the object of a verb. To simplify patterns like **F8**, all verbs in the sentence are lemmatized with *morpha* (Minnen et al., 2001) before the pattern matching begins.

F1 *it* VERB ADJ *that*
F2 *it* VERB ADJ
   what/which/where/whether/why/how
F3 *it* VERB ADJ *to*
F4 *it* VERB *that*
F5 *it* VERB *what/which/where/whether/why/how*
F6 *it* VERB *to*
F7 *it* ADJ *that/to*
F8 *it* be/seem as if
F9 *it* VERB COMMA

For each item above, the feature consists of the distance (number of tokens) between *it* and the end of the match (the right bracket such *that* or *to*). By using the distance as the feature, it is possible to avoid specifying a cutoff point for the end of a match. The memory-based learning algorithm can adapt to the training data. As discussed in Section 2.1, extraposition is often lexically triggered, so the specific verbs and adjectives in the sentence are important for its classification. For this reason, it is necessary to include information about the surrounding verbs and adjectives. The nearby full verbs

(as opposed to auxiliary and modal verbs) are likely to give the most information, so we add features for the immediately preceding full verb (for **F7**), the following full verb (for **F1-F6**), and the following adjective (for **F1-3,7**). The verbs were lemmatized with *morpha* and added as features along with the following adjective.

F10 lemma of immediately preceding full verb
F11 lemma of following full verb within current sentence
F12 following adjective within current sentence

5.2 Cleft *it*

Two patterns are used for cleft *it*:

F13 *it be who/which/that*
F14 *it who/which/that*

As mentioned in the previous section, all verbs in the sentence are lemmatized before matching. Likewise, these features are the distance between *it* and the right bracket. Feature **F14** is used to match a cleft *it* in a phrase with inverted word order.

5.3 Weather/Condition/Time/Place *it*

Ideally, the possible weather predicates could be learned automatically from the following verbs, adjectives, and nouns, but the list is so open that it is better in practice to specify a fixed list. The weather/time/place/condition predicates were taken from the training set and put into a fixed list. Some generalizations were made (e.g., adding the names of all months, weekdays, and seasons), but the list contains mainly the words found in the training set. There are 46 words in the list. As Denber mentioned, WordNet could be used to extend this list. A feature is added for the distance to the nearest weather token.

The following verb lemma feature (**F10**) added for extrapositional *it* is the lemma of the following full verb, but in many cases the verb following weather *it* is the verb *be*, so we also added a binary feature for whether the following verb is *be*.

F15 distance to nearest weather token
F16 whether the following verb is *be*

5.4 Idiomatic *it*

Idioms can be identified by fixed patterns. All verbs in the sentence are lemmatized and the following patterns, all found as idioms in our training data, are used:

if/when it come to	*pull it off*
as it happen	*fall to it*
call it a NOUN	*ask for it*
on the face of it	*be it not for*
have it not been for	*like it or not*

Short idiom patterns such as "cool it" and "watch it" were found to overgeneralize, so only idioms including at least three words were used. A binary feature was added for whether an idiom pattern was matched for the given instance of *it* (**F17**). In addition, two common fixed patterns were included as a separate feature:

it be ... time
it be ... my/X's turn

F17 whether an idiom pattern was matched
F18 whether an additional fixed pattern was matched

5.5 Additional Restrictions

There are a few additional restrictions on the pattern matches involving length and punctuation. The first restriction is on the distance between the instance of *it* and the right bracket (*that*, *to*, *who*, etc.). On the basis of their corpus, Paice and Husk decided that the right bracket could be at most 27 words away from *it*. Instead of choosing a fixed distance, features based on pattern matches are the distance (number of tokens) between *it* and the right bracket.

The system looks for a pattern match between *it* and the end of the sentence. The end of a sentence is considered to be punctuation matching any of the following: . ; : ? !)] . (Right parenthesis or bracket is only included if a matching left parenthesis or bracket has not been found before it.) If there is anything in paired parentheses in the remainder of the sentence, it is omitted. Quotes are not consistent indicators of a break in a sentence, so they are ignored. If the end of a sentence is not located within 50 tokens, the sentence is truncated at that point and the system looks for the patterns within those tokens.

As Paice and Husk noted, the presence of a single comma or dash between *it* and the right bracket is a good sign that the right bracket is not relevant to whether the instance of *it* is non-referential. When there are either zero or two or more commas or dashes it is difficult to come to any conclusion without more information. Therefore, when the total comma count or total dash count between *it* and the right bracket is one, the pattern match is ignored.

Additionally, unless *it* occurs in an idiom, it is also never the object of a preposition, so there is an additional feature for whether *it* is preceded by a preposition.

F19 whether the previous word is a preposition

Finally, the single preceding and five following simplified part-of-speech tags were also included. The part-of-speech tags were simplified to their first character in the C7 tagset, adverb (R) and negative (X) words were ignored, and only the first instance in a sequence of tokens of the same simplified type (e.g., the first of two consecutive verbs) was included in the set of following tags.

F20-25 surrounding POS tags, simplified

6 Results

Training and testing data were generated from our corpus using the the 25 features described in the previous section. Given Evans's success and the limited amount of training data, we chose to also use TiMBL's k-nearest neighbor algorithm (IB1). In TiMBL, the distance metric can be calculated in a number of ways for each feature. The numeric features use the numeric metric and the remaining features (lemmas, POS tags) use the default overlap metric. Best performance is achieved with gain ratio weighting and the consideration of 2 nearest distances (neighbors). Because of overlap in the features for various types of non-referential *it* and sparse data for cleft, weather, and idiomatic *it*, all types of non-referential *it* were considered at the same time and the output was a binary classification of each instance of *it* as referential or non-referential. The results for our TiMBL classifier (MBL) are shown in Table 7 alongside our results using a decision tree algorithm (DT, described below) and the results from our replication of Evans

	Our MBL Classifier	Our DT Classifier	Repl. of Evans
Accuracy	88%	81%	76%
Precision	82%	82%	57%
Recall	71%	42%	60%

Table 7: Results

Extrapositional	81%
Cleft	45%
Weather	57%
Idiomatic	60%
Referential	94%

Table 8: Recall by Type for MBL Classifier

(2001). All three systems were trained and evaluated with the same data.

All three systems perform a binary classification of each instance of *it* as referential or non-referential, but each instance of non-referential *it* was additionally tagged for type, so the recall for each type can be calculated. The recall by type can been seen in Table 8 for our MBL system. Given that the memory-based learning algorithm is using previously seen instances to classify new ones, it makes sense that the most frequent types have the highest recall. As mentioned in Section 2.2, clefts can be difficult to identify.

Decision tree algorithms seem suited to this kind of task and have been used previously, but C4.5 (Quinlan, 1993) decision tree algorithm did not perform as well as TiMBL on our data, compare the TiMBL results (MBL) with the C4.5 results (DT) in Table 7. This may be because the verb and adjective lemma features (**F10**-**F12**) had hundreds of possible values and were not as useful in a decision tree as in the memory-based learning algorithm.

With the addition of more relevant, generalized grammatical patterns, the precision and accuracy have increased significantly, but the same cannot be said for recall. Because many of the patterns are designed to match specific function words as the right bracket, cases where the right bracket is omitted (e.g., extraposed clauses with no overt complementizers, truncated clefts, clefts with reduced relative clauses) are difficult to match. Other problematic cases include sentences with a lot of intervening

material between *it* and the right bracket or simple idioms which cannot be easily differentiated. The results for cleft, weather, and idiomatic *it* may also be due in part to sparse data. When only 2% of the instances of *it* are of a certain type, there are fewer than one hundred training instances, and it can be difficult for the memory-based learning method to be very successful.

7   Conclusion

The accurate classification of *it* as referential or non-referential is important for natural language tasks such as reference resolution (Ng and Cardie, 2002). Through an examination of the types of constructions containing non-referential *it*, we are able to develop a set of detailed grammatical patterns associated with non-referential *it*. In previous rule-based systems, word lists were created for the verbs and adjectives which often occur in these patterns. Such a system can be limited because it is unable to adapt to new texts, but the basic grammatical patterns are still reasonably consistent indicators of non-referential *it*. Given a POS-tagged corpus, the relevant linguistic patterns can be generalized over part-of-speech tags, reducing the dependence on brittle word lists. A machine learning algorithm is able to adapt to new texts and new words, but it is less able to generalize about the linguistic patterns from a small training set. To be able to use our knowledge of relevant linguistic patterns without having to specify lists of words as indicators of certain types of *it*, we developed a machine learning system which incorporates the relevant patterns as features alongside part-of-speech and lexical information. Two short lists are still used to help identify weather *it* and a few idioms. The k-nearest neighbors algorithm from the Tilburg Memory Based Learner is used with 25 features and achieved 88% accuracy, 82% precision, and 71% recall for the binary classification of *it* as referential or non-referential.

Our classifier outperforms previous systems in both accuracy and precision, but recall is still a problem. Many instances of non-referential *it* are difficult to identify because typical clues such as complementizers and relative pronouns can be omitted. Because of this, subordinate and relative clauses cannot be consistently identified given only a POS-tagged corpus. Improvements could be made in the future by integrating chunking or parsing into the pattern-matching features used in the system. This would help in identifying extrapositional and cleft *it*. Knowledge about context beyond the sentence level will be needed to accurately identify certain types of cleft, weather, and idiomatic constructions.

References

L. Burnard, 1995. *Users reference guide for the British National Corpus*. Oxford.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2003. TiMBL: Tilburg Memory Based Learner, version 5.0, Reference Guide. ILK Technical Report 03-10. Technical report.

Michel Denber. 1998. Automatic resolution of anaphora in English. Technical report, Imaging Science Divison, Eastman Kodak Co.

Richard Evans. 2001. Applying machine learning toward an automatic classification of It. *Literary and Linguistic Computing*, 16(1):45 – 57.

Rodney D. Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge.

Shalom Lappin and Herbert J. Leass. 1994. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4):535–561.

Guido Minnen, John Caroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

Vincent Ng and Claire Cardie. 2002. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*.

C. D. Paice and G. D. Husk. 1987. Towards an automatic recognition of anaphoric features in English text; the impersonal pronoun 'it'. *Computer Speech and Language*, 2:109 – 132.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.