to appear in *Proc. 5th Intl. Workshop on Computational Semantics,* Tilburg, 2003

Inducing criteria for mass noun lexical mappings using the Cyc KB, and its extension to WordNet

Tom O'Hara[‡], Nancy Salay, Michael Witbrock, Dave Schneider,
Bjørn Aldag, Stefano Bertolo, Kathy Panton, Fritz Lehmann[*],
Jon Curtis, Matt Smith, David Baxter, and Peter Wagner

Cycorp, Inc. Austin, TX 78731; www.cyc.com
{tom, nancy, witbrock, daves, aldag, bertolo,
panton, fritz, jonc, msmith, baxter, peter}@cyc.com

[‡]Computer Science Department, New Mexico State University
Las Cruces, NM 88001; www.cs.nmsu.edu/~tomohara
tomohara@cs.nmsu.edu

December 6, 2002

Abstract

This paper presents an automatic approach for learning semantic criteria for the *mass versus count noun* distinction by induction over the lexical mappings contained in the Cyc knowledge base. This produces accurate results (89.5%) using a decision tree that only incorporates semantic features (i.e., Cyc ontological types). Comparable results (86.9%) are obtained using OpenCyc, the publicly available version of Cyc. For broader applicability, the mass noun criteria using Cyc are converted into criteria using WordNet, preserving the general accuracy (86.3%).

1 Introduction

In semantic lexicons where the underlying concept is represented separately from the word or phrase being defined, a *lexical mapping* is used to establish the connection between the two [ON95, BD99]. For convenience, the term *lexicalize* will refer to the process of producing these mappings, which are referred to as *lexicalizations*.[1] Deciding whether the headword in a phrase should be

[*]Now at Legicode, Inc.

[1]The term *lexicalization* is used in a broader sense than that traditionally used in grammatical literature: "fossilized" words (i.e., no longer morphologically decomposable [HP02]).

lexicalized as a mass noun is not as straightforward as it might seem. There are guidelines available in traditional grammar texts, as well as the more technical linguistics literature. But these mainly cover high level categories, such as substances, the prototypical category for mass nouns, and concrete objects, the prototypical category for count nouns. However, for lower-level categories the distinctions are not so clear, especially when the same headword occurs in different types of contexts. For example, "source code" represents a mass noun usage, whereas "postal code" would be a count noun usage.

In addition, sometimes the same word will be a mass noun in some contexts and a count noun in others, depending on the underlying concept. For example, "anthrax" will be a mass noun when referring to the bacteria, but it will be a count noun when referring to the resulting skin lesions (e.g., "anthraces"). There has been much work on the coercion of count nouns into mass nouns (and vice versa), such as the 'grinding rule' [BCL95], a special case of which covers animal terms becoming mass nouns when referring to the food (e.g., "Let's have pig tonight."). However, there has been little work on determining whether terms should be lexicalized via mass nouns or count nouns. The work here illustrates how this can be done by learning a decision tree based on the ontological types of the underlying concept. Thus, it relies only upon semantic criteria.

Motivation for this comes in the context of building a large-scale knowledge base (KB), namely Cyc [Len95]. Traditionally at Cycorp, there has been a split in the knowledge engineering, with the domain knowledge being entered separately from the lexical knowledge. The reason for this is that the knowledge engineers might not be familiar with the linguistic considerations necessary for performing the mappings accurately. They also might not be familiar with all the lexicalization conventions to allow for consistent lexical knowledge entry. To alleviate this bottleneck, the linguistic criteria can be inferred from the knowledge base, exploiting the large number of previous decisions made by lexical knowledge engineers regarding speech part selection.

This problem is not just restricted to computational lexicons such as the Cyc English lexicon. As an example, web search engines often provide suggestions as follow-ups to particular queries, albeit with infelicitous English. For instance, in its suggestions, *Yahoo!* exhibits some knowledge of the mass versus count noun distinction, but it seems to have incomplete coverage:[2]

Query	*Suggestion*
"luggage"	over 50 Luggage listings on Yahoo! Auctions
"table"	over 100 Tables on Yahoo! Auctions
"desk"	over 100 Desk on Yahoo! Auctions

This shows that 'luggage' is correctly given as a mass noun, and 'table' is correctly pluralized for the suggestions; however, 'desk' is incorrectly left singular. The technique presented here can help such systems produce better wording for their automatically generated web pages. To provide for more general applicability, an extension is included where the Cyc terms used in the criteria

[2]The searches were produced using the basic search option via www.yahoo.com.

are mapped into WordNet terms (i.e. synsets [Mil90]).[3] Note that unlike an approach that relies on a list of known mass nouns, such as from a learner's dictionary, this can handle novel headwords provided that the underlying types are known (e.g., closest WordNet synset).

After an overview of the Cyc knowledge base in the next section, Section 3 discusses the approach taken to inferring the part of speech for lexicalizations, along with the classification results. Section 4 then covers the extension to WordNet. This is followed by a comparison to related work in Section 5.

2 Cyc knowledge base

The Cyc knowledge base is a vast repository of commonsense knowledge that has been in development for over 15 years [Len95], containing over 120,000 concepts and a million assertions that interrelate them.[4] At its highest level, the KB consists of an ontology describing how the world is generally conceptualized by human beings (e.g., objects versus stuff). At the other extreme, it contains a grab bag of miscellaneous facts useful for particular applications, such as web searching, but not necessarily representative of commonsense reasoning (e.g., that "Dubya" refers to *President George W. Bush*). In between resides the area of the KB most associated with commonsense reasoning, such as relating to various human activities. In addition, the KB includes a broad-coverage English lexicon mapping words and phrases to terms throughout the KB.

2.1 Ontology

Part of the ontology is a taxonomy of concepts[5] that are partially ordered via two *hierarchical* relations: *isa* (i.e., is-instance-of) and *genls* (i.e., has-generalization). These correspond to Cruse's [Cru86] *relation of dominance* and specify the type definition for a concept. In addition, there are a variety of *non-hierarchical* relations providing additional information, such as attributes, relations to other concepts, and usage restrictions.

Figure 1 shows the type definition for *PhysicalDevice*,[6] a prototypical denotatum term for count nouns. Concept names in Cyc generally are self-explanatory, so descriptions are not included unless relevant to the discussion. However, Table 1 describes some of the common types terms used in Cyc; these are used later in the experiments. Note that *ExistingObjectType* is unintuitively a specialization of *TemporalStuffType*.

[3] WordNet is a popular resource that makes explicit the lexical relations typically contained in dictionaries and thesauri. An online version is available at www.cogsci.princeton.edu/~wn, along with the database files and documentation.

[4] These figures and the results discussed later are based on Cyc KB version 576 and system version 1.2577. See www.cyc.com/publications.html for detailed documentation on the KB.

[5] Atomic terms in the KB are called *constants*; there are also non-atomic terms (e.g., *(LeftFn Brain)*), for which the type definitions are inferred automatically.

[6] Unless otherwise noted, all examples are taken from OpenCyc version 0.7 (KB version 567 and system version 1.2594). An online version is available at www.opencyc.org/public_servers.

Collection: **PhysicalDevice**

Mt: ArtifactGVocabularyMt
isa: ExistingObjectType
genls: Artifact ComplexPhysicalObject SolidTangibleProduct

Mt: ProductGMt
isa: ProductType

Figure 1: Type definition for *PhysicalDevice*, a prototypical denotatum term for count noun mappings. (*G-Mt* indicates a general microtheory.)

The type definition of *PhysicalDevice* indicates that it is a collection (i.e., category) that is a specialization of *Artifact*, etc. In addition, it is an instance of *ExistingObjectType*, which is typical for terms referred to by count nouns. Note that the 'Mt' labels refer to microtheories, which is the way that knowledge is organized in Cyc to facilitate contextual inferences as well as to account for the needs of different applications [Guh90].

Figure 2 shows the type definition for *Water*, a prototypical denotation for mass nouns. Although the type information for *Water* does not convey any properties that would suggest a mass noun lexicalization, the assertions under *ChemicalCompoundType*, its type, do clearly suggest this type of mapping. However, since *ChemicalCompoundType* is a specialization of *TangibleStuffCompositionType*, *Water* is an instance of the latter. Thus a mass noun lexicalization is appropriate. This illustrates that the decision tree for the mass noun distinction needs to consider inherited types, along with immediate type assertions.

2.2 English lexicon

In Cyc, natural language lexicons are integrated directly into the KB [BD99]. There are several natural language lexicons in the KB, kept separate via microtheories, but the English lexicon is the only full-scale one. The mapping from phrases to concepts is done through a variety of lexical assertions. These fall into two broad categories, corresponding to proper names and common noun phrases. Proper name assertions map strings to individuals in the KB (i.e., non-collections). For example,

(nameString Taiwan-RepublicOfChina "Nationalist China")

A *denotational assertion* maps a phrase into a concept, usually a collection. The phrase is specified via a lexical word unit (i.e., lexeme concept) with optional string modifiers. The part of speech is specified via the one of Cyc's *SpeechPart* constants. Syntactic information, such as the wordform variants and their speech parts, is stored under the constant for the word unit. For example, *Device-TheWord*, the word unit for 'device', just has a single syntactic mapping since the plural form is inferred:

Concept	Description
Thing	the "universal collection"
PartiallyIntangible	things having an intangible part
Intangible	things that are not physical
Individual	things that are neither sets nor collections
IntangibleIndividual	wholly intangible individuals
MathematicalThing	atemporal, non-spatial, mathematical things
Collection	natural kinds (not mathematical sets)
PartiallyIntangible-Individual	individual with some intangible component
TemporalThing	things with temporal extent or location
SpatialThing	things with a spatial extent or location
ObjectType	differentiated entities (i.e., having 'parts' that are not also instances of the collection)
StuffType	undifferentiated entities (i.e., every 'part' is also instance of the collection)
TemporalStuffType	same as *StuffType* with respect to time slices
ExistingObjectType	temporally stuff-like (*TemporalStuffType*) but spatially object-like (*ObjectType*)
ExistingStuffType	temporally stuff-like (*TemporalStuffType*) as well as spatially stuff-like (*StuffType*)

Table 1: Examples of Cyc ontological types. These are used as features in the experiments.

Collection: **ChemicalCompoundType**

Mt: UniversalVocabularyMt
isa: AtemporalNecessarilyEssentialCollectionType CollectionType

Mt: BaseKB
isa: SiblingDisjointCollectionType CollectionType PublicConstant

Mt: NaivePhysicsVocabularyMt
isa: SecondOrderCollection

Mt: UniversalVocabularyMt
genls: TangibleStuffCompositionType

Collection: **Water**

Mt: NaivePhysicsVocabularyMt
isa: ChemicalCompoundType
genls: Oxide

Figure 2: Type definition for *Water*, a prototypical denotatum term for mass noun mappings, including the definition for *ChemicalCompoundType*, its type.

	Usage	
Predicate	OpenCyc	Cyc
denotation	3218	16589
compoundString	330	1958
multiWordString	1252	23670
headMedialString	192	871
total	4992	43088

Table 2: Denotational predicate usage in the Cyc English lexicon. This excludes microtheories for non-standard lexicalizations (e.g., *ComputereseLexicalMt*).

Constant: Device-TheWord
Mt: GeneralEnglishMt
isa: EnglishWord
posForms: CountNoun
singular: "device"

The simplest type of denotational mapping associates a particular sense of a word with a concept via the *denotation* predicate (i.e., relation type). For example,

(denotation Device-TheWord CountNoun 0 PhysicalDevice)

This indicates that sense 0 of the count noun 'device' refers to *PhysicalDevice* via the associated wordforms "device" and "devices".

To account for phrasal mappings, three additional predicates are used, depending on the location of the headword in the phrase. These are *compoundString*, *headMedialString*, and *multiWordString* for phrases with the headword at the beginning, the middle, and the end, respectively. For example,

(multiWordString ("women's") Wear-TheWord MassNoun WomensClothing)

This states that "women's wear" refers to *WomensClothing*. Since the lexical mapping is through a mass noun usage of the word 'wear', there are no variants of the phrase.

Table 2 shows the frequency of the various predicates used in the denotational assertions, excluding lexicalizations that involve technical, informal or slang terms.[7] Of these, 9,739 have a *MassNoun* part of speech for the headword, compared to 20,936 for *CountNoun*. This subset of the denotational assertions forms the basis of the training data used in the mass versus count noun classifier, as discussed later.

[7]Cyc was adapted for use in the Hotbot web search engine and thus recognizes many colorful mass terms (e.g., "farm sex").

2.3 OpenCyc

In the spring of 2002, Cycorp released a portion of the KB as an open source resource called OpenCyc.[8] It include over 8,000 atomic concepts[9] and more than 99,000 assertions, which is roughly 7% of the entire KB. Over time, Cycorp intends to release larger parts of the KB, to promote the integration of Cyc into intelligent applications. In addition, there are plans for a research version of the system closer in scope to the full KB.

Such a public resource as OpenCyc will be valuable for natural language processing as well as for other areas of artificial intelligence. The work here shows that the lexical information represented in the KB is not only useful for knowledge-based systems but can also be adapted for use in applications employing machine learning.

3 Inference of lexicalization part of speech

3.1 General approach

Our method of inferring the part of speech for noun lexicalizations is to apply machine learning techniques over the lexical mappings from English words or phrases to Cyc terms. For each target denotatum term, the corresponding types and generalizations are extracted from the ontology. This includes terms for which the denotatum term is an instance or specialization, either explicitly asserted or inferable via transitivity. For simplicity, these are referred to as *ancestor terms*. The association between the lexicalization parts of speech and the common ancestor terms forms the basis for the criteria used in the mass-count classifier.[10]

There are several possibilities in mapping this information into a feature vector for use in machine learning algorithms. The most direct method is to have a binary feature for each possible ancestor term, but this requires thousands of features. To prune the list of potential features, frequency considerations can be applied, such as taking the most frequent terms that occur in type definition assertions. Alternatively, the training data can be analyzed to see which reference terms are most correlated with the classifications.

For simplicity, the frequency approach is used here. The most-frequent 256 atomic terms are selected, excluding internal constants flagged with the *quotedCollection* predicate (e.g., *PublicConstant*); half of these terms are taked from the *isa* assertions, and the other half from the *genls* assertions. These are referred to as the *reference terms*. For instance, *ObjectType* is a type for 21,042 of the denotation terms (out of 43,088 cases), compared to 19,643 for *StuffType*.

[8] Information on OpenCyc is available at www.opencyc.org.

[9] This excludes the 1,000+ functional concepts, such as *(JuvenileFn Dog)*, which is used in place of *Puppy*, but includes linguistic constants (e.g., *Device-TheWord*).

[10] This same process can be applied to the full set of speech part category values; but syntactic features would be necessary for accurate results.

These occur at ranks 10 and 11, so they are both included. In contrast, *Hand-Tool* occurs only 226 times as a generalization term at rank 443, so it is pruned. Some of the top terms after pruning were shown previously in Table 1, along with informal descriptions.

Given a training instance, such as a denotation from a word unit into a specific Cyc concept using a particular *SpeechPart* (e.g., *MassNoun* or a *Count-Noun*), the feature specification is derived by determining all the ancestor terms of the denotatum term and converting this into a vector of occurrence indicators, one indicator per reference term. The part of speech serves as the classification variable. For example, consider the mapping of "heat production" to *HeatProductionProcess*.

> (multiWordString ("heat") Produce-TheWord *MassNoun HeatProductionProcess*)

The type definition follows along with some of the ancestor terms inferred via transitivity (as given in the Cyc KB Browser).

> Collection: *HeatProductionProcess*
> Mt: NaivePhysicsVocabularyMt
> isa: TemporalStuffType DefaultDisjointScriptType
> genls: Emission
>
> (isa HeatProductionProcess ?ARG2)
> 32 answers for ?ARG2 :
> Collection ... StuffType ... TemporalStuffType Thing
>
> (genls HeatProductionProcess ?ARG2)
> 22 answers for ?ARG2 :
> Emission EnergyTransferEvent Event Event-Localized
> GeneralizedTransfer ... Thing TransferOut Translocation

It turns out that all of these except for *EnergyTransferEvent* are in the reference list. Therefore, the corresponding feature vector would have 1's in the 49 slots corresponding to the unique reference terms and 0's in the other 207 slots, along with *MassNoun* for the classification value.

The example illustrates that some of the reference terms are not very relevant to the classification at hand (e.g., *Thing*). Advanced techniques could be used to address this, such as that used for collocation selection in word-sense disambiguation based on conditional probability [WMB98]. This is not done here, as it complicates the training process without significantly improving performance. The result is a table containing 30,675 feature vectors that forms the training data. Standard machine learning algorithms can then be used to induce the mass noun lexicalization criteria.

3.2 Sample criteria

We use decision trees for this classification. Part of the motivation is that the result is readily interpretable and can be incorporated directly by knowledge-

based applications. Decision trees are induced in a process that recursively splits the training examples based on the feature that partitions the current set of examples to maximize the information gain [WF99]. This is commonly done by selecting the feature that minimizes the entropy of the distribution (i.e., yields least uniform distribution). Because the complete decision tree is over 300 lines long, just a few fragments are shown to give an idea of the criteria being considered in the count-mass classification.

(1) if ObjectType and Event and CreationEvent then
    if AnimalActivity then
        CountNoun
    else
        MassNoun

This fragment indicates that creation events are generally lexicalized via count noun mappings when they represent animal activities. Otherwise, mass noun lexicalizations are used. An example of a concept inheriting from *AnimalActivity* is *MakingSomething*, with the count term "creation". One not inheriting from *AnimalActivity* is *PhysicalSynthesis*, with the mass term "physical synthesis."

(2) if (not ObjectType) and (not Relation) and Agent-Generic then
    MassNoun

    if (not ObjectType) and Relation then
        CountNoun

The second rule fragment indicates that if both *ObjectType* and *Relation* are not ancestor terms for a concept, then the reference will use mass nouns for concepts that inherit from *Agent-Generic*. An example of this is *Dissatisfied*, referred to as "dissatisfaction". The notion of generic agents might seem odd here, but emotional states in Cyc are restricted to agents. For concepts that are not typed as *ObjectType* but are typed as *Relation*, the reference will use count nouns. For example, any *UnitOfMeasure*, a specialization of *Relation*, is lexicalized using a count noun (e.g., "meter").

3.3 Results

Table 3 shows the results of 10-fold cross validation for the mass-count classification.[11] This was produced using the J48 algorithm in the Weka machine learning package [WF99].[12] This shows that the system achieves an accuracy of 88.7%, an improvement of 21.3 percentage points over the baseline of always

[11] 10-fold cross validation involves randomly partitioning the data into 10 parts, each of which serves as the test data in one trial (with the rest use for the training data). The trials are averaged to give the overall accuracy [MS99, WF99].

[12] Weka is freely available via www.cs.waikato.ac.nz/~ml/weka/index.html.

	OpenCyc	Cyc
Instances	3395	30675
Entropy	0.74	0.90
Baseline	79.2	68.2
Accuracy	86.9	*89.5*

Table 3: Mass-count classification over Cyc lexical mappings and using Cyc reference terms as features. *Instances* refers to size of the training data. *Baseline* selects most frequent case. *Accuracy* is average in the 10-fold cross validation.

selecting the most frequent case.[13] The OpenCyc version of the classifier also performs well. This suggests that sufficient data is already available in OpenCyc to allow for good approximations for such classifications.

Note that these results are obtained strictly via semantic features (i.e., Cyc's ontological types). The use of headword morphological features should improve the performance. For instance, English has quite a few suffixes indicative of mass noun usages [QGLS85], such as '-age', '-ery', and '-ism' (e.g., "baggage", "slavery", and "idealism"). Work is underway at Cycorp to make the relations among words more explicit, which should allow for further improvements.

4 Extension to WordNet

The mass noun criteria based on the full Cyc KB requires access to the KB to be useful for incorporation in applications. The full KB is proprietary except for certain research purposes, so access to it might be difficult. However, the criteria induced over the Cyc KB can be carried over into WordNet by taking advantage of the WordNet mapping in the KB (covering a subset of WordNet version 1.6). In effect, this augments the WordNet lexicon with mass noun indicators, making it easier for applications such as *Yahoo!* to account for the distinction.

The Cyc-to-WordNet mapping includes over 8,000 of the synsets, with emphasis on the higher-level Cyc concepts. The mapping could be applied either to the final decision tree or to the feature table prior to classification. The latter is preferable, because the decision tree induction can then account for overly general mappings along with gaps in the mappings.

A separate classifier based on WordNet synsets is produced as follows: Each of the Cyc reference term features is replaced by a feature for the corresponding reference synset. Each of these binary features indicates whether the target denotatum synset is a specialization of the reference synset:

⟨target-synset, *has-ancestor-hypernym*, reference-synset⟩

Correspondence is established by first checking for an assertion directly linking the Cyc reference term to a WordNet synset. If that fails, there is a check for

[13] Accuracy here is the same as precision in information retrieval. As is often the case with decision trees, exactly one answer is provided for all instances, so recall equals precision.

	OpenCyc	Cyc
Instances	3395	30675
Entropy	0.74	0.90
Unmapped accuracy	86.9	89.5
Baseline	79.2	68.2
Mapped accuracy	85.3	*86.3*

Table 4: Mass-count classification over Cyc lexical mappings using reference term features mapped into WordNet. *Baseline* selects most frequent case. *Unmapped accuracy* refers to results shown earlier. *Mapped accuracy* incorporates the WordNet mappings prior to training and classification (average of 10 trials).

a linkage from one of the reference term's generalizations into WordNet. In cases where there are no such synsets, the feature will not be used. In cases where several reference terms correspond to the same synset, the features will be conflated.

Given the 256 reference terms used for the Cyc-based results (shown in Table 3), the process to establish correspondences yields 70 distinct features (due to 62 deletions and 124 conflations). Table 4 shows the results, indicating an accuracy of 86.3% in mass-noun classification, which is close to that when using the original features.

The following is a simple fragment from the resulting decision tree:

(3) if N03875475 then {*color, coloring*}
    if N04496504 then {*kind, sort, form, variety*}
        CountNoun
    else
        MassNoun

This shows that color terms are generally mass nouns unless referring to kinds of colors (e.g., different pigments). In terms of WordNet, since the corresponding synsets are disjoint (i.e., not related via a common hypernym), this entails that the mass noun lexicalization will always be preferred. In Cyc, the count noun usage only applies when concepts are lexicalized via multi-word phrases headed by "color" (e.g., *HumanSkinColor* as "skin color"). These concepts are not represented in WordNet, so this does not produce any conflicts.

5 Related work

We are unaware of other approaches to the automatic determination of the mass-count distinction, using either statistical or traditional knowledge-based frameworks. However, there has been much work on the interpretation of mass terms in the formal semantics literature, especially with regard to logical form representation and quantifier scoping issues (e.g., [Løn89, PS84]). Bunt [Bun85]

presents an account of mass term interpretation via his *Ensemble Theory*, which is built up around *part-whole* relations, using an alternative set theory axiomatization which uses the subset operation ($\subseteq$) rather than the instance operation ($\in$) as a primitive. Note that his and related work address a different aspect of mass term interpretation, namely how the terms are interpreted in context. For example, he shows how the ensemble approach facilitates modeling of the modification of mass nouns (e.g., "the snow in the garden"), avoiding problems that occur with traditional set-theoretical approaches. In contrast, we address the creation of lexical mappings of mass terms into concepts, which can be viewed as precompiling mass noun preferences into the lexicon. In fact, this could serve as input into Bunt's process for mass noun interpretation.

Quirk et al. [QGLS85] provide rough guidelines for whether nouns will be mass nouns or count nouns based on the type of the denotation term. For example, count nouns refer to individual countable entities whereas mass nouns refer to undifferentiated masses (or continua). Huddleston and Pullum [HP02] provide similar guidelines but do provide more details on the differences involved. For instance, the main criteria for mass terms is that the concepts be perceived as being inherently unbounded. This accounts for heterogeneous collections that are given mass noun lexicalizations (e.g., "luggage").

Gillon [Gil99] suggests that the mass-count distinction can be determined using the notion of aggregation. In particular, the denotations of count nouns refer to sets made up of elements that are the minimal aggregates for which the term applies. In contrast, the denotations of mass nouns refer to a singleton set consisting of the maximal aggregate for which the term applies. For example, "bird" denotes the set of all the individual avian animals, whereas "fowl" denotes the set whose sole member is the aggregation of the avian animals (loosely speaking, the same set viewed collectively). This is an elegant account of the distinction, but it does not admit of an immediate decision procedure. For example, most collections in Cyc are not specifically typed as to whether they are undifferentiated aggregates or not. There is the *StuffType* versus *ObjectType* distinction, but this is generally applied to the high-level collections. In addition, due to multiple inheritance, there are many lower-level collections that are both typed as *StuffType* and *ObjectType*.

There has been more work on 'conversions' from count noun usages to mass noun and vice versa. For example, the grinding rule mentioned earlier [BCL95] converts a count noun interpretation for an individuated object into a mass noun interpretation for a substance. The animal grinding rule is a special case for when count nouns that refer to animals are used as mass nouns to refer to the animal used as food. Gillon [Gil99] generalizes this and similar cases to a rule that converts a count noun usage for any object to a mass noun usage referring to an aggregate part of the object (e.g., meat in the case of the animal grinding rule).

6 Conclusion

This paper shows that an accurate decision procedure (89.5%) for determining mass-count distinction in lexicalizations can be induced from the lexical mappings in the Cyc KB. This relies solely on semantic information, in particular Cyc's ontological types, illustrating the degree to which this distinction can be made without syntactic considerations. In practice, it should be augmented with other criteria such as ones based on the morphology of the headword for the mappings. Although the main approach relies on Cyc's conceptual distinctions, the results can be incorporated in other applications via the WordNet mapping.

Future work will investigate additional features for the mass-count lexicalization classifier, in particular features based on morphology. In addition, we will look into how nouns in context might be classified as to being count nouns or mass nouns. We will also investigate both extensions in the context of general speech-part classification for lexicalizations. This would complement existing part-of-speech taggers by allowing for more detailed tag types.

In closing, given the recent release of OpenCyc, we encourage others to investigate how information in the Cyc knowledge base can be exploited to infer criteria for other interesting phenomena related to natural language processing and other intelligent applications.

Acknowledgements

Many staff members at Cycorp have helped directly or indirectly with the lexical work discussed here, including Doug Foxvog, Keith Goolsbey, Zelal Güngördü, Robert Kahlert, Charles Klein, Doug Lenat, Kim Loika, Daniel Mahler, Matt Olken, Karen Pittman, and Jennifer Sullivan. In addition, Steve Reed and John DeOlivera have been very helpful with their work for OpenCyc. The lexicon work at Cycorp has been supported in part by grants from NIST, DARPA (e.g., RKF), and ARDA (e.g., AQUAINT).

The first author is formerly of Cycorp and currently supported by a generous GAANN fellowship from the Department of Education. The work was facilitated by computing resources at NMSU made possible through MII Grants EIA-9810732 and EIA-0220590.

References

[BCL95] Ted Briscoe, Ann Copestake, and Alex Lascarides. Blocking. In P. Saint-Dizier and E. Viegas, editors, *Computational Lexical Semantics*, pages 273–302. Cambridge University Press, Cambridge, 1995.

[BD99] Kathy J. Burns and Anthony B. Davis. Building and maintaining a semantically adequate lexicon using Cyc. In Viegas [Vie99], pages 121–143.

[Bun85] Harry C. Bunt. *Mass terms and model-theoretic semantics*. Cambridge University Press, Cambridge, 1985.

[Cru86] D. A. Cruse. *Lexical Semantics*. Cambridge University Press, Cambridge, 1986.

[Gil99] Brendan S. Gillon. The lexical semantics of English count and mass nouns. In Viegas [Vie99], pages 19–37.

[Guh90] R. V. Guha. Micro-theories and contexts in Cyc. Technical Report ACT-CYC-129-90, Microelectronics and Computer Technology Corporation, Austin, TX, 1990.

[HP02] Rodney Huddleston and Geoffrey K. Pullum. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, 2002.

[Len95] D. B. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11), 1995.

[Løn89] Jan Tore Lønning. Computational semantics of mass terms. In *Proc. of the 4th EACL*, pages 205–211, Manchester, UK, 1989.

[Mil90] G. Miller. Special issue on WordNet. *International Journal of Lexicography*, 3(4), 1990.

[MS99] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, 1999.

[ON95] B. Onyshkevych and S. Nirenburg. A lexicon for knowledge-based MT. *Machine Translation*, 10(2):5–57, 1995. Special Issue on Building Lexicons for MT.

[PS84] Francis Jeffry Pelletier and Lenhart K. Schubert. Two theories for computing the logical form of mass expressions. In *Proc. Coling-84*, 1984.

[QGLS85] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvick. *A Comprehensive Grammar of the English Language*. Longman, 1985.

[Vie99] Evelyn Viegas, editor. *The Breadth and Depth of Semantic Lexicons*. Kluwer, 1999.

[WF99] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.

[WMB98] Janyce Wiebe, Kenneth McKeever, and Rebecca Bruce. Mapping collocational properties into machine learning features. In *Proc. 6th Workshop on Very Large Corpora (WVLC-98)*, pages 225–233, 1998.