Bracketing Compound Nouns for Logic Form Derivation

Vasile Rus  
Department of Computer Science  
Southern Methodist University  
Dallas, TX 75275-0122  
{vasile@engr.smu.edu}

Dan I. Moldovan  
Department of Computer Science  
University of Texas at Dallas  
Dallas, TX  
{moldovan@utdallas.edu}

Orest Bolohan  
Department of Computer Science  
University of Texas at Dallas  
Dallas, TX  
{orest@utdallas.edu}

We present here a method for bracketing coordinated compound nouns in the larger context of deriving first order logic forms. The method consists of two phases: (1) detection - extract candidate compound nouns from a parse tree and (2) interpretation - provide a bracketing solution for each candidate. Bracketing performance on 525 coordinations is 87.42%. This compares favorably with a 52.35% baseline when a default split by the coordination is applied.

Introduction

WordNet (Miller 1995) can be viewed as a rich source of world knowledge structured based on lexico-semantic relations among concepts represented as sets of words that have same meaning (synsets). Each synset has a gloss or a small textual definition and few examples attached to it. We aim at transforming WordNet glosses into a computational representation that enables reasoning mechanisms. In this paper we address the issue of bracketing coordinated compound nouns.

The logic form that we use is first order logic and includes syntactic information in the form of positional arguments. It was first introduced by Hobbs (Hobbs 1986) and enhanced by Harabagiu, Miller and Moldovan (Harabagiu, Miller, & Moldovan 1999). For each content word a predicate is generated. Verbs, prepositions and conjunctions have the role of linking predicates describing relations among entities/events denoted by their arguments. Adverbs and adjectives have a modifier role and share the same argument with their modifiers.

Our approach to derive the logic form is to use structural information available in a syntactic tree (Moldovan & Rus 2001). We use an in house implementation (done by Mihai Surdeanu) of Collins’ statistical model for parsing (Collins 1997). For the case of compound nouns the parser does not help us at all. Collins’ model interprets coordinated compound nouns as base NPs and does not provide any bracketing/structure inside. A flat treatment of coordinated compound nouns similar to a regular coordinated base NP (split the noun phrase by the conjunction as illustrated in is the most straightforward approach and the least accurate. This flat treatment is considered baseline when comparing the performance of different approaches.

Related Work

Our work resembles efforts to extract lexical information from machine readable dictionaries (MRD), as LDOCE (Longman Dictionary of Contemporary English) or Webster’s 2nd International Dictionary (W2). Different parsing methods of definitions were used: pattern-matching (Chodorow, Byrd, & Heidorn 1985), specially constructed definition parsers (Wilks, Slator, & Guthrie 1996) or broad coverage parsers (Richardson, Dolan, & Vanderwende 1998) (ISI 1998). All those efforts were limited to extracting genus terms, unlabeled or labeled relations or to build taxonomies (ISI 1998). We parse WordNet glosses to generate logic representations that enable reasoning mechanisms.

Problem Description

Compound nouns are sequences of nouns that together have an enhanced meaning, sometimes different, as compared to individual nouns. An example is goat hair where the two nouns refer to the hair of goat which is different from the general concept of hair, though related, and different from the goat concept. A coordinated compound noun is one or more compound nouns linked via a coordinated conjunction as in goat hair and camel hair. The usage of coordinated compound nouns features a language laziness in the form of an ellipsis: goat and camel hair.

Figure 1 shows two possible bracketing alternatives and their corresponding logic forms for the previous coordinated compound noun: goat and camel hair. There is a larger range of alternatives to choose from when modifiers are involved, e.g. common house and field crickets from gloss of Acheta:n#1 leads to three different bracketing possibilities. The paper presents a method to bracket coordinated compound nouns for further deriving logic forms.

When bracketing coordinated compound nouns types of problems encountered are manifold. Those errors can be classified in two distinct classes: detection of coordinated compound nouns and interpretation of coordinated compound nouns.

Detection errors

A noun phrase having a sequence of tags of the form NN[PS] CC NN[PS] NN[PS] after the elimination of determiners and modifiers does not always lead to a coordinated compound noun ([]) means alternative extensions to capture
In addition to detection errors there are interpretation errors that appear after correctly detecting coordinated compound nouns. Interpretation errors are strongly linked to bracketing. For example, the coordinated compound noun hair/NN or/CC finger/NN nails/NNS from gloss (scissors for cutting hair or finger nails) of clipper:n#4 should be bracketed as (NP

\(\text{NN hair} \ (\text{CC or} \ (\text{NNS finger_nails}))\) and its logic form would be: \([\text{or}(x_1, x_2, x_3) \ & \ \text{hair}(x_2) \ & \ \text{finger_nail}(x_3)]\) while peach/NN or/CC almond/NN trees/NNS in gloss (willow of the western United States with leaves like those of peach or almond trees) of peach_leaf/NN willow:n#1 should be bracketed as (NP (NP (NN peach) (CC or) (NN almond)) (NP (NNS trees))) and its logic form would be: \([\text{or}(x_1, x_2, x_3) \ & \ \text{peach}(x_2) \ & \ \text{almond}(x_3) \ & \ \text{nn}(x_4, x_5) \ & \ \text{tree}(x_5)].\)

Table 1 provides alternative bracketings and the corresponding logic form for both examples. Our task is to select a bracketing that would lead to the correct logic form.

**Bracketing Coordinated Compound Nouns**

To derive the logic form for coordinated compound nouns we designed our own algorithm that takes advantage of POS tagging, structural information from parser, semantic sense of nouns involved and type of coordinated conjunctions. It consists of two steps: (i) provide a bracketing for the coordinated compound noun (ii) apply the rule-based approach presented in (Moldovan & Rus 2001). In turn, the bracketing consists of two steps: (1) detection and (2) interpretation.

**Coordinated Compound Nouns Detection**

In the following this simplified frame: \(n_1 \ cc \ n_2 \ h\) is used where \(n_1\) is the first noun, \(n_2\) is the second noun and \(h\) is the head noun in a coordinated base NP of the form NN CC NN NN. As we saw previously, there are many factors that influence the accuracy of detecting coordinated compound nouns. We provide solutions for each of them.

**POS Tagging**

The first major source of errors for detecting coordinated compound nouns is part of speech (POS) tagging. The POS tagging module tags glosses by applying a two levels voting scheme between Brill's rule based tagger (Brill 1992) and MXPOST statistical tagger (Ratnaparkhi 1996) at the first level (Mihalcea & Moldovan 2001), and between Brill's and WordNet syntactic category information on a second level (Rus 2001). On top of this two levels voting scheme we developed a WordNet-biased Brill tagger, called wnBrill, by automatically deriving new contextual rules from a corpus of 3,000 correctly tagged glosses built by hand. 40 new contex-
Coordinated rules were added and 6 were eliminated from the original set due to their bad behaviour characterized by providing more errors than corrections. An example of rule that was eliminated is VBZ NNS SURROUND TAG NN which indicates a change from VBZ onto NNS when NN is a surrounding tag. The set of eliminated rules consists of rules too specific to the training set. Training the tagger on different corpora and then check which rules withstand can be a way to detect the most generally applicable contextual rules for POS tagging. When tested on 1,000 new glosses (again manually tagged) after voting and applying the specialized wnBrill tagger the overall accuracy was 99.14%. The bare tagger leads to 95.63% and thus our method results in a 3.51% reduction in error rate. There is one interesting aspect of tagging glosses in the larger context of parsing and deriving logic forms that we want to point out. When a single tag in one gloss is erroneous the entire parse process of that gloss is error-prone. From this perspective we propose to use a newly tagging measure $P_{TC}$, called exact sentence accuracy, which is defined as the number of glosses with all words correctly tagged divided by the total number of glosses attempted. This new measure reflects the upper limit of how well a parser can perform when using a specific tagger as a preprocessor. Using this measure the Brill’s bare performance is 61.08%. Using voting and a wnBrill tagger we managed to improve the exact sentence measure to 90.59% or a 29.51% improvement.

### Compound Concept Detection
Unrecognized common compound concepts may also lead to false coordinated compound noun detection. For example in gloss of water_gas:n#1, hydrogen/NN and/CC carbon/NN monoxide/NN can be misinterpreted as a coordinated compound noun unless one identifies carbon_monoxide as a single concept in WordNet. On the other hand for the base NP bond/NN or/CC stock/NN shares/NNS from gloss of flotation:n#2, stock_shares can be considered as a single concept (Moldovan & Girju 2001) in some specific domains, whereas generally speaking is not. Our option would be to go for the general case that can be easily tested by checking whether stock_shares has an entry in WordNet or not. We form pairs of concepts $n_1$ and $n_2$ and look them up in WordNet. This test works only inside base NPs and not for every sequence of two or more words. Here is a counter example: covering/VBG designed/VBN to/T0 be/VB worn/VBN on/IN a/DT person/NN’s/POS body/NN which is the gloss of clothing:n#1 we were worn on is identified as a WordNet concept and it should not be considered for the case of this gloss. We treat common compound nouns separately from NEs because of future extensions of this work in which we envision using some dictionaries of NEs to compensate for the incompleteness of WordNet in this aspect. Extensions of WordNet with other common compound concepts it is a more challenging task and the reader can find more details in (Moldovan & Girju 2001).

### Coordinated Compound Nouns Interpretation
At this point the possible compound concepts are identified and the main task of this phase is to provide a correct brack-

<table>
<thead>
<tr>
<th>Coordination</th>
<th>Bracketing</th>
<th>Logic Form</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1. hair/NN or/CC</td>
<td>(NP (NP (NN hair)) (CC or) (NP (NN finger) (NNS nails)))</td>
<td>[or(x1,x2,x3) &amp; hair:(x2) &amp; nn(x3,x4,x5) &amp; finger(x4) &amp; nail(x5)]</td>
<td>No</td>
</tr>
<tr>
<td>1.2. hair/NN or/CC</td>
<td>(NP (NP (NN hair) (CC or) (NP (NNS nails)))</td>
<td>[nn(x1,x2,x3) &amp; hair:(x4) &amp; or(x2,x4,x5) finger(x5) &amp; nail(x3)]</td>
<td>No</td>
</tr>
<tr>
<td>1.3. hair/NN or/CC</td>
<td>(NP (NN hair) (CC or) (NNS nails))</td>
<td>[or(x1,x2,x3) &amp; hair:(x2) &amp; finger_nails(x3)]</td>
<td>Yes</td>
</tr>
<tr>
<td>2.1. peach/NN or/CC</td>
<td>(NP (NP (NN peach)) (CC or) (NP (NN almond) (NNS trees)))</td>
<td>[peach(x1) &amp; or(x2,x1,x3) &amp; nn(x3,x4,x5) &amp; almond(x4) &amp; tree(x5)]</td>
<td>No</td>
</tr>
<tr>
<td>2.2. peach/NN or/CC</td>
<td>(NP (NP (NN peach) (CC or) (NP (NN almond) (NNs trees)))</td>
<td>[or(x1,x2,x3) &amp; peach(x2) &amp; almond(x3) &amp; nn(x4,x1,x5) &amp; tree(x5)]</td>
<td>Yes</td>
</tr>
<tr>
<td>2.3. peach/NN or/CC</td>
<td>(NP (NP (NN peach)) (CC or) (NP (NN almond) (NNS trees)))</td>
<td>[or(x1,x2,x3) &amp; peach(x2) &amp; almond_tree(x3)]</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 1: Alternative bracketing solutions and their logic forms
The presence of the same word on both sides of the coordination shows the user’s desire to refer to two different sets: one described by \( n \) and another formed by instances of a compound concept modified by \( n \).

For example gloss (the industry that makes steel and steel products) of concept steel\_industry:n#1 contains the coordination steel and steel products. **HEURISTIC 1** solves this coordination by bracketing it as (NP (NN steel)) (CC and) (NP (NN steel) (NN products))).

**HEURISTIC 2**: if \( ((n_1 \ h \ IS \ in \ WordNet) \ AND \ (n_2 \ h \ IS \ in \ WordNet)) \) then bracket \( (\langle n_1 \ cc \ n_2 \rangle \ h) \)

**Rationale.** If \( n_1 \ h \) and \( n_2 \ h \) are concepts in WordNet and some author wants to coordinate them he will most likely use a compact form for the coordination for language efficiency reasons (fewer words means less effort to express himself), especially when the head of the two compound nouns is the same.

For example, the coordination North and Central America from the gloss (tropical marine bivalve found chiefly off eastern Asia and Pacific coast of North and Central America) attached to pearl\_oyster:n#1 has North America and Central America as concepts that are found in WordNet and thus the bracketing is (NP (NP (NN North) (CC and) (Central))) (NP (NN America))).

The next heuristic deals with the situation when \( n_1 \) and \( n_2 \) are siblings of common concept \( c \). It requires disambiguation information which it is not a hardly accepted assumption keeping in mind that the glosses will soon be disambiguated as part of the XWN project. For the test cases we manually tagged the sense of nouns in coordinations.

**HEURISTIC 3**: if \( ((n_1 \ IS \ sibling \ of \ e) \ AND \ (n_2 \ IS \ sibling \ of \ e)) \) then bracket \( (\langle n_1 \ cc \ n_2 \rangle \ h) \)

**Rationale.** When two satellite nouns in a coordination are siblings of same concept the author intended to express two different subsets of the common concept and in order to emphasize these relation places them in a coordination with a shared head.

An example solved by this heuristic is tomato and potato plants from the gloss of concept tomato\_hornworm:n#1: (large green white-striped hawkmoth larva that feeds on tomato and potato plants). The bracketing provided is (NP (NP (NN tomato) (CC and) (NN potato)) (NP (NN plants))).

When none of the previous heuristic triggers on a specific coordination the default approach of splitting by the coordination is applied.

**Modifiers**

All previous discussions were focused on coordinations that did not contain modifiers. There are many coordinations (29.94% of all coordinated compound nouns found on noun hierarchy) that contain modifiers and here we face the issue of what concept they modify. For a frame such as \( jj^* n_1 \ cc \ n_2 \ h \) the modifier \( jj \) may be attached to \( n_1 \), to \( (n_1 \ cc \ n_2) \), or \( (n_1 \ cc \ n_2 \ h) \) (* means one or more occurrences of modifiers). An example is that of an/DT annual/JJ school/NN or/CC university/NN reunion/NN from the gloss (an annual school or university reunion for graduate) of concept homecoming:n#1, where annual should modify both school reunion and university reunion. Sometimes the presence of a modifier provides us with a hint for bracketing: large in gloss (extremely active cylindrical squid with short strong arms and large rhombic terminal fins) of concept ommastrephes:n#1, acting as a modifier after the coordination indicates that terminal fins is a noun phrase of its own and the bracketing should be: (NP (NP (JJ short injustic arms) (CC and) (NP (JJ large) (JJ rhombic) (NN terminal)) (NN fins))). We designed several heuristics that attach the modifiers to their modifiers.

**HEURISTIC 4**: if (exists \( jj_1 \) AND \( jj_2 \)) then bracket \( ((jj_1 \ n_1 \ cc \ (jj_2 \ n_2 \ h)) \)

This heuristics says that when there is a modifier in front of the first noun \( n_1 \) and at the same time immediately after the coordination \( cc \), the coordination separates two different concepts which do not share the head. The rationale is that when the speaker adds a modifier to both \( n_1 \) and \( n_2 \) he wants to emphasize a distinction between the two.

**HEURISTIC 5**: if ((exists \( jj_1 \) AND (\( jj_2 \ IS NOT) AND (cc = 'or')) then bracket \( ((jj_1 \ n_1 \ cc \ (jj_2 \ n_2 \ nn)) \)

This second heuristic simply says that when there is no modifier after the coordination and the coordination is or then \( j_1 \) is attached to the whole coordinated compound noun. The rationale is that when only one modifier is present it is most likely that the speaker wants to attach the modifier to both nouns. The presence of or comes from an empirical observation that this heuristic does not apply when an and is used. The example given in the first paragraph of this section (from gloss of homecoming:n#1) is covered by this heuristic.

**HEURISTIC 6**: if ((NOT exists \( jj_1 \) AND (exists \( jj_2 \)) then bracket \( (n_1 \ cc \ (jj_2 \ n_2 \ h)) \)

The rationale for this heuristic is simple: whenever the speaker places a modifier before two following nouns the modifier most likely attaches to the compound concept of the two nouns. An example is mid\_waters and deep\_slope\_waters from gloss of dory:n#2. The right bracketing for this example is (NP (NP (NN mid\_waters)) (CC and) (NP (JJ deep) (NN slope)) (NN waters))).

After the bracketing is done a set of transformation rules are applied to trees in order to obtain logic forms (see (Moldovan & Rus 2001))(Rus 2001)).

**Experiments and Results**

All our experiments were performed on a set of 522 coordinated compound nouns extracted from the WordNet noun hierarchy: 298 simple coordination (without modifiers) and 224 of them contain modifiers.

First we focus on the 298 simple coordinations. The definitions are extracted from glosses, tagged using Brill’s tagger, expanded to full sentences to minimize parser’s errors.
and then parsed. Base NPs are detected that contain this sequence of tags: NN[PS] CC NN[PS] NN[PS]. The overall detection precision, defined as the number of correctly identified coordinated compound nouns over the total number, is $P_D = 55.07\%$. This simple detection approach is poor in terms of precision.

Table 2 shows the distribution of the main source of errors at detection. Most errors are due to coordinated Named Entities, followed by POS tagging errors and compound concepts identification. Parse errors are insignificant (1%). After applying the solutions proposed for each source of errors the overall precision jumps to $P_D = 94.46\%$. The taggers disagree on 152 cases out of which 78 or 51.31% are automatically assigned using new set of rules, 38 are passed to WordNet agreement and 16 (out of 38) or 10.52% (42% of the 38 passed cases) are automatically corrected. The user intervenes in 47 cases. The NE module recognizes NE in 62 coordinations: 26 are simple, 31 are coordinated and 5 are all-way. The classification module misclassifies only one coordination: Central/NNP and/CC South/NNP Africa/NNP of gloss of Bantu:n#1: (a member of any of a large number of linguistically related peoples of Central and South Africa). South Africa is a concept in WordNet while in this particular gloss it refers to the region of southern Africa. Common coordinated compound nouns are identified based on semantic information tagged manually.

Detected coordinations are pipelined into the interpretation module which applies the three heuristics presented previously (the default approach of splitting by the coordination is applied to 44 cases). Using a straightforward approach such as providing a bracketing by splitting along the coordination: $(n_1) \text{ cc } (n_2) \text{ h}$ leads to a bracketing precision of 38.48%. Using the heuristics the bracketing precision jumps to 83.52%. The notable difference between the precision of the two approaches is mainly explained by the large number of coordinated proper compound nouns such as South and Central America.

For coordinations with modifiers we tested the performance of attaching the modifiers on 224 cases. The very first heuristic applied to 120 of them, the second on 58 and the third heuristic on 46. The performance of each heuristic is given in the fourth column of Table 3.

On the 525 cases the overall performance is 87.42% measured as number of correctly bracketed coordinations divided by all coordinations. The default approach of splitting by coordinations leads to 52.35% precision when applied to all 525 test cases.

<table>
<thead>
<tr>
<th>Type</th>
<th>Percentage</th>
<th>Eliminated</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS Tagging</td>
<td>9.45%</td>
<td>6.78%</td>
</tr>
<tr>
<td>Compound Concepts</td>
<td>8.78%</td>
<td>4.91%</td>
</tr>
<tr>
<td>Name Entity</td>
<td>25%</td>
<td>24.59%</td>
</tr>
<tr>
<td>Parse</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>Total</td>
<td>44.93%</td>
<td>36.28%</td>
</tr>
</tbody>
</table>

**Table 3: Performance of the proposed heuristics**

<table>
<thead>
<tr>
<th>Name</th>
<th>Tempted</th>
<th>Solved</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heuristic 1</td>
<td>6</td>
<td>6</td>
<td>100%</td>
</tr>
<tr>
<td>Heuristic 2</td>
<td>112</td>
<td>103</td>
<td>91.96%</td>
</tr>
<tr>
<td>Heuristic 3</td>
<td>61</td>
<td>53</td>
<td>86.88%</td>
</tr>
<tr>
<td>Heuristic 4</td>
<td>120</td>
<td>114</td>
<td>95%</td>
</tr>
<tr>
<td>Heuristic 5</td>
<td>58</td>
<td>51</td>
<td>87.93%</td>
</tr>
<tr>
<td>Heuristic 6</td>
<td>46</td>
<td>46</td>
<td>100%</td>
</tr>
</tbody>
</table>

**Conclusions**

We presented here a set of heuristics to deal with the problem of bracketing coordinated compound nouns in the larger context of deriving logic forms. We experimented with 525 candidate coordinations on which 87.42% bracketing precision was achieved.

**References**


Rus, V. 2001. Logic Form Derivation for WordNet Glosses. Southern Methodist University, Dallas, TX.