A Noun Phrase Parser of English

Atro Voutilainen
Helsinki

Abstract

An accurate rule-based noun phrase parser of English is described. Special attention is given to the linguistic description. A report on a performance test concludes the paper.

1. Introduction

1.1 Motivation.

A noun phrase parser is useful for several purposes, e.g. for index term generation in an information retrieval application; for the extraction of collocational knowledge from large corpora for the development of computational tools for language analysis; for providing a shallow but accurately analysed input for a more ambitious parsing system; for the discovery of translation units, and so on. Actually, the present noun phrase parser is already used in a noun phrase extractor called *NPtool* (Voutilainen 1993).

1.2. Constraint Grammar.

The present system is based on the Constraint Grammar framework originally proposed by Karlsson (1990). A few characteristics of this framework are in order.

- The linguistic representation is based on surface-oriented morphosyntactic tags that can encode dependency-oriented functional relations between words.

- Parsing is reductionistic. All conventional analyses are provided as alternatives to each word by a context-free lookup mechanism, typically a morphological analyser. The parser itself seeks to discard all and only the contextually illegitimate alternative readings. What 'survives' is the parse.

- The system is modular and sequential. For instance, a grammar for the resolution of morphological (or part-of-speech) ambiguities is applied, before a syntactic module is used to introduce and then resolve syntactic ambiguities.

Proceedings of NODALIDA 1993, pages 301-310

- The parsing description is based on linguistic generalisations rather than probabilities. The hand-written rules, or **constraints** are validated against representative corpora to ensure their factuality. Also heuristic constraints can be used for resolving remaining ambiguities.

- Morphological analysis is based on two-level descriptions (Koskenniemi 1983). Large lexicons and informative morphosyntactic descriptions are used to represent the core vocabulary of the language. Words not recognised by the morphological analyser are processed with a very reliable heuristic analyser.

- Parsing is carried out with linear-precedence constraints that discard morphological or syntactic readings in illegitimate contexts. Typically, a constraint expresses a partial generalisation about the language.

1.3. System architecture

A typical analyser in this framework also thepresent one employs the following sequentially applied components:

1. Preprocessing
2. Morphological analysis
3. Morphological heuristics
4. Morphological disambiguation
   - 4a. Grammar-based constraints
   - 4b. Heuristic constraints
5. Lookup of alternative syntactic tags
6. Syntactic disambiguation
   - 6a. Grammar-based constraints
   - 6b. Heuristic constraints

Descriptions pertaining to modules 1–4 are directly adopted from the ENGCG description, written by Voutilainen, Heikkil and Anttila, and documented in Voutilainen, Heikkil and Anttila (1992), Karlsson, Voutilainen, Heikkil and Anttila (Eds.) (forthcoming). Here, only the barest characteristics of modules1–4 in effect, a part-of-speech tagger are mentioned. The reader is referred to Karlsson et al. (forthcoming) for further details and justifications.

- The preprocessor recognises sentence boundaries, idioms and compounds. The ENGTWOL morphological analyser employs a 56,000-entry lexicon and a morphosyntactic description based on Quirk et al. (1985). Some 93–98 % of all word-form tokens in running text become recognised. 'Morphological heuristics' is a rule-based module that assigns ENGTWOL-style analyses to those words not recognised by ENGTWOL itself. About 99.5 % of these heuristic

predictions are correct. Ambiguity in English is a nontrivial problem: on an average, the ENGTWOL analyser furnishes two alternative morphological readings for each word.

- The morphological disambiguator applies a grammar with a set of 1,100 'grammar-based' and another set of 200 heuristic constraints. After the combined application of these 1,300 constraints, 96–98 % of all word form tokens in the text are morphologically unambiguous, while at least 99.6 % of all word-form tokens retain the correct morphological reading. These figures apply to standard non-fiction English. The accuracy may decrease somewhat if the text is colloquial, fiction, dialectal or otherwise non-standard. – To my knowledge, this precision/recall ratio is by far the best in the field.

2. Parsing scheme

The ENGCG description also contains a syntactic grammar based on a parsing scheme of some 30 function tags. The somewhat unoptimal recall and precision of the syntactic description on the one hand, and the observation that the parsing scheme was unnecessarily delicate for some of the applications mentioned above, on the other, motivate a more ascetic parsing scheme. I have designed as new syntactic parsing scheme with only seven function tags that capitalise on the opposition between noun phrases and other categories on the one hand, and between heads and modifiers, on the other. Next, the tags are presented.

- *@V* represents auxiliary and main verbs as well as the infinitive marker to in both finite and non-finite constructions. For instance:

  She should/@V know/@V what to/@V do/@V

- *@NH* represents nominal heads, especially nouns, pronouns, numerals, abbreviations and *-ing*-forms. Note that of adjectival categories, only those with the morphological feature *<Nominal>*, e.g. *English*, are granted the *@NH* status: all other adjectives (and *-ed*-forms) are regarded as too unconventional nominal heads to be granted this status in thepresent description. An example:

  The English/@NH may like the unconventional

- *@>N* represents determiners and premodifiers of nominals (the angle-bracket '>' indicates the direction in which the head is to be found). The head is the following nominal with the tag *@NH*, or a premodifier in between. For instance, consider the analysis of *fat* in *fat butcher's wife*:

fat/@>N butcher's/@>N wife/@NH

The annotation accounts for both of the following bracketings:

[[fat butcher's] wife]
[[fat [butcher's wife]

Our tag notation leaves implicit certain structurally unresolvable distinctions in order to maximise on the accuracy of the parser. For instance, on structural criteria it is impossible to decide whether the butcher or his wife is fat in this case. To avoid the introduction of certain other types of semantic or higher-level distinctions, the tag @>N represents not only what are conventionally described as determiners and premodifiers: also non-final parts of compounds as well as titles are furnished with this tag, e.g. *Mr./@>N Jones* and *Big/@>N Board.*

- *@N<* represents prepositional phrases that unambiguously postmodify a preceding nominal head. Such unambiguously postmodifying constructions are typically of two types: (i) in the absence of certain verbs like 'accuse', postnominal *of*-phrases and (ii) preverbal NP-PP sequences, e.g.

  The man in/@<N the moon had a glass of/@N< ale.

  Structure-based resolution of the attachment ambiguities of prepositionalphrases that are preceded by a verb and immediately by a noun phrase is often very difficult or impossible (Quirk et al. 1985). To maximise on the informativeness of the syntactic analysis, the present description capitalises on the unambiguously resolvable 'easy' cases without paying the penalty of introducing systematic unresolvable ambiguity in the hardcases. It is, however, still quite easy to identify the inherently ambiguous cases, if necessary: they are prepositional phrases tagged as *@AH*, and they are preceded by a nominal head.

Currently the description does not account for other types of postmodifier, e.g. postmodifying adjectives, numerals, other nominals, or clausal constructions. Clausal constructions are ignored because their accurate treatment presupposes effective control of clause-level information (or clause boundaries), which is hard to employ in the present description. Besides, postmodifying clauses would probably be marginal for some applications, at least for index term generation.

- @AH represents adjectival heads, adverbials of various kinds, adverbs (also intensifiers), and also those of the prepositional phrases that cannot be dependably analysed either as an adverbial or as a postmodifier. For example:

  There/@AH have always/@AH been extremely/@AH many people around/@AH.

  Note in passing that ed-forms occurring after the primary verbs 'be' and 'have' are generally analysed as main verbs rather than as *@AH*'s, to which status they could in principle be ranked as potential (adjectival) subject complements. A uniform analysis one way or the other (*@V* vs. *@AH*) is not harmful here because neither category qualifies as a nounphrase in the present application. Besides, the ambiguity due to the subject complement and main verb reading in this type of configuration tends to be unresolvable on structural, and often even on any other, criteria, so the present uniform analysis saves us from some (structurally) unmotivated ambiguity.

- *@CC* and *@CS* are familiar from the ENGCG description: the former represents co-ordinating conjunctions, and the latter represents subordinating conjunctions. For example:

  Either/@CC you or/@CC I will go if/@CC necessary.

Finally, a short sample output of the parser is in order:

```
("<*the>"
   ("the" <*> <Def> DET CENTRAL ART SG/PL (@>N)))
("<inlet>"
   ("inlet" N NOM SG (@>N @NH)))
("<and>"
   ("and" CC (@CC)))
("<exhaust>"
   ("exhaust" N NOM SG (@>N)))
("<manifolds>"
   ("manifold" N NOM PL (@NH)))
("<are>"
   ("be" <SV> <SVC/N> <SVC/A> V PRES -SG1,3 VFIN (@V)))
("<mounted>"
   ("mount" <SVO> <SV> <P/on> PCP2 (@V)))
("<on>"
   ("on" PREP (@AH)))
("<opposite>"
   ("opposite" <Nominal> A ABS (@>N)))
("<sides>"
```

```
   ("side" N NOM PL (@NH)))
("<of>"
   ("of" PREP (@N<)))
("<the>"
   ("the" <Def> DET CENTRAL ART SG/PL (@>N)))
("<cylinder>"
   ("cylinder" N NOM SG (@>N)))
("<head>"
   ("head" N NOM SG/PL (@NH)))
("<$.>")
```

Here *inlet* remains ambiguous due to the modifier and head functions because of a coordination ambiguity.

3. About the parsing grammar

The syntactic grammar contains some 120 syntactic constraints, some 50 of which are heuristic. Like the morphological disambiguation constraints, these constraints are essentially negative partial linear-precedence definitions of the syntactic categories. The present grammar is a partial expression of four general grammar statements:

1. *Part of speech determines the order of determiners and modifiers.*
2. *Only likes coordinate.*
3. *A determiner or a modifier has a head.*
4. *An auxiliary is followed by a main verb.*

We will give only one illustration of how these general statements can be expressed as constraints. A partial paraphrase of the statement *Part of speech determines the order of determiners and modifiers*: 'A premodifying noun occurs closest to its head'. In other words, premodifiers from other parts of speech do not immediately follow a premodifying noun. Therefore, a noun in the nominative immediately followed by an adjective is not a premodifier.Thus a constraint would discard the @>N tag of *Harry* in the following sample sentence, where Harry is directly followed by an unambiguous adjective:

```
("<*is>"
   ("be" <SVC/N> <SVC/A> V PRES SG3 (@V)))
("<*harry>"
   ("harry" <Proper> N NOM SG (@NH @>N)))
("<foolish>"
   ("foolish" A ABS (@AH)))
("<$?>")
```

We require that the noun in question is a nominative because premodifying nouns in the genitive can occur also before adjectival premodifiers; witness *Harry's* in *Harry's foolish self.*

Regarding the heuristic elements in the grammar, the main strategy is to prefer the premodifier function over head function. The underlying heuristicis that a noun phrase is not directly followed by another unless there is an explicit noun phrase edge – e.g. a determiner or a genitive in between.

4. A test run

The parser was tested against a text collection new to the system. In all, 3,600 words from newspapers, detective stories, technical abstracts and book reviews were analysed. Some of the texts contained characteristics from spoken language and fiction, so the corpus can be considered a somewhat hard test bench for the system.

Of all words, 93.5 % became syntactically unambiguous, and 99.15 % of all words retained the most appropriate syntactic reading, i.e. 31 contextually appropriate readings were discarded. (A little over 97 % of all words became **morphologically** unambiguous; also heuristic constraints were used.) Of these 31 errors, 18 were due to the syntactic constraints; 11 were due to disambiguation constraints, and 2 were due to the ENGTWOL lexicon. Some observations about the misanalyses are in order.

- Errors tend to co-occur. In the following sentence fragments, four contextually legitimate infinitives were discarded by the morphological disambiguator (the misanalysed word is indicated with a slash, followed by the discarded feature.

  ..either to enhance (boost/INF or increase/INF) or to suppress (dampen/INF or decrease/INF) other nodes' activation.

  One of the constraints discards an infinitive if to the left, there is another unambiguous infinitive, and in between, there is neither a coordinating conjunction nor another infinitive marker (e.g. *to* or a modal auxiliary). Parenthetical expressions of this kind were ignored in the grammar, so both *boost* and *dampen* lost their infinitive readings, retaining some other verb readings. The infinitive readings of *increase* and *decrease* were lost as a domino effect: a constraint about coordination forbade a sequence consisting of a non-infinitive verb coordinating conjunction infinitive.

- Generally, a pronoun does not take a determiner or a premodifier. Heuristic constraints capitalise on this, resulting in the following misanalyses:

  ..that is, the same/DET ones should underlie..
  ..are general/@>N ones.

  The determiner reading of *same* as well as the premodifier reading of *general* is discarded. These errors are actually quite easy to correct: *one* is an untypical pronoun in that it quite often takes a determiner or a premodifier. Correcting the relevant constraints presupposes the addition of another context condition that in effect functions as a brake: whenever the pronoun happens to be a form of *one*, a preceding determiner or premodifier reading is left intact.

- In the following cases, the morphological disambiguator lost two noun readings:

  Peanut-butter tan/N.
  Expensive gold watch/N.

  Non-clausal utterances that are not marked as such (e.g. with a heading code) are known to be problematic for the present description, based on the assumption that an utterance ending with a fullstop or a question mark or an exclamation mark is a sentence with at least one finite verb. In the above cases, the finite verb readings of *tan* and *watch* were selected because no other finite verb candidates were available in the 'sentence'.

- Above, it was mentioned that some heuristic syntactic constraints prefer the premodifier function over the head function. A couple of misanalyses resulted:

  ..the relationship/@NH Ashdown had confessed..
  During the same campaign/@NH Tory politicians told..

- Multi-word adjectives turned out to be the most fatal single error source for the syntactic constraints:

  ..might not be language/@>N specific.
  ..error/@>N prone..
  A Cell/@>N Organized Raster Display for Line Drawings <ENDTITLE> " Attribute/ Based File Organization..

  There is a constraint that discards the premodifier function tag of a noun if the following word is an adjective (or a non-finite *ed*-form).

This generalisation misses adjectives consisting of a noun-adjective sequence, e.g. *language specific*. This leak in the grammar can be mended to some extent at least by imposing lexical context-conditions that licence a premodifying noun in front of certain adjectives or non-finite *ed*-forms such as *specific* or *based*, both of which seem to be quite productive in the formation of multi-word adjectives. A representative collection of these adjectives can be extracted from large ENGCG-tagged corpora relatively easily.

Overall, it seems to me that relatively few of the misanalyses are elementary from the point of view of higher-level syntactic generalisations; in terms of lexical knowledge, these errors can often be quite easily anticipated. For instance, a better version of the grammar may still reject premodifying nouns in general in case the following word is an adjective but a limited class of known exceptions, such as *specific*, can be accounted for by imposing further lexical context conditions. The more accurate the present description becomes, the more lexico-grammatically oriented it is likely to be.

These observations seem to bear on a more general question about how lexical information can be employed in structural analysis, such as part-of-speech disambiguation. One view held in the literature has, roughly speaking, been to identify using structural information with grammar-based methods, and using lexical information (as lexical preferences) with statistical methods (see e.g. Church 1992; Church and Mercer 1993). Our observation is that information about lexis certainly is a useful addition to more general structural information, and, more importantly, lexical information can also be employed in a grammar-based system, such as the present reductionistic one. Furthermore, the superior recall/precision ratio of the present system suggests that a rule (or knowledge) based use of lexical information, in conjunction with more general structural information, may be preferable over using lexical information in the form of probabilities.

5. Technical information

The ENGTWOL morphological analyser uses the two-level program by Kimmo Koskenniemi and Lingsoft, Inc. The latest version of the Constraint Grammar parser was written by Pasi Tapanainen. Also several Unix utilities are used in the present prototype. On a Sun SPARCstation 10/30, the whole system from preprocessing through syntax analyses some 400 words per second. Some optimisation efforts would be worthwhile; at present, much of the processing time is taken by very simple operations that have not been implemented effectively. The hardest problem of parsing with a large grammar has already been

addressed quite satisfactorily: disambiguation and syntactic analysis together can be carried out at a speed of more than 1,000 words per second.

The system will become available. Contact the author for further details, e.g. by email to Atro.Voutilainen@Helsinki.FI.

References

Church, K. 1992. *Current Practice in Part of Speech Tagging and Suggestions for the Future.* In Simmons (ed.) 1992. *Sbornik praci: In Honor of Henry Kucera.* Michigan Slavic Studies.

Church, K. and R. Mercer. 1993. *Introduction to the Special Issue on Computational Linguistics Using Large Corpora.* COMPUTATIONAL LINGUISTICS, Vol. 19, Number 1.

Karlsson, F. 1990. *Constraint Grammar as a framework for parsing running text.* In Karlgren, H. (Ed.) COLING–90. *Papers presented to the 13th International Conference on Computational Linguistics*, Vol. 3. Helsinki, Finland.

Karlsson, F., A. Voutilainen, J. Heikkil and A. Anttila (eds.). (Forthcoming). *Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text.* Mouton deGruyter.

Koskenniemi, K. 1983. *Two-Level Morphology: a General Computational Model for Word-Form Recognition and Production.* Publication No. 11, Department of General Linguistics, University of Helsinki.

Quirk, Randolph, S. Greenbaum, G. Leech and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language.* Longman, London.

Voutilainen, A. 1993. *NPtool, a detector of English noun phrases.* In *Proceedings of the Workshop on Very Large Corpora, June 22, 1993.* Ohio State University, Ohio, USA.

Voutilainen, A., J. Heikkil and A. Anttila. 1992. *Constraint grammar of English. A Performance-Oriented Introduction.* Publications No. 21, Department of General Linguistics, University of Helsinki.