English Verb Analyzer: Identifying tense, voice, aspect, sense and grammatical meaning in context for pedagogic purposes

John Blake
University of Aizu
Aizuwakamatsu
Japan
jblake@u-aizu.ac.jp

Abstract

This paper describes the development of an online pedagogic tool that helps learners of English notice and understand how verbs are used. The English Verb Analyzer automatically identifies tense and grammatical meaning in context. When users submit texts, all finite verb groups are classified and colorized according to one of twelve grammatical tenses (e.g. present perfect simple). On click, users can access more information on the verb group. The information currently incorporated includes grammatical meaning in context, voice, verb sense and lexical meaning. The grammatical meaning is predicted based on the combination of verb tense, verb sense and verb class. This prototype improves on an earlier version by using a decision-tree approach to identify match tense and harnessing the supersense functionality in WordNet to assign grammatical meaning in context more accurately.

1 Introduction

Learners of English as an additional language frequently struggle with grammatical tenses. The lack of one-to-one mapping between their mother tongues and English is one cause of misunderstanding and misuse. This is particularly so for speakers of languages that are more linguistically distant, such as Chinese, Japanese and Arabic.

Language may be acquired through extensive exposure over many years, but learners of English in educational establishments in many countries have limited exposure to English-medium materials and little (if any) daily interaction with users of English. Their grasp of grammatical tenses tends to be gleaned from prescriptive rules in textbooks.

Learners of English may not realize that "I'm cutting my finger" implies repetition but "I'm holding my finger" does not. The English Verb Analyzer aims to help learners see how lexical aspect and grammatical tense combine to create meaning. Learner awareness of tense usage in any text is raised by colorizing the verbs by tense. According to the noticing hypothesis, noticing is a necessary precursor to learning (Schmidt, 2012). Apart from an earlier prototype, this is the only online tool that can identify tense and provide additional information about verb groups on demand.

Tense identification is non-trivial due to *inter alia* the polysemic nature of verbs, difficulties in disambiguation, and accurately identifying parts of speech (POS). A fully functional tense identifier has the potential to create a sea change in the teaching and learning of the tense system for learners of English as an additional language.

This paper is organized as follows. Section 2 introduces the twelve grammatical tenses, and provides an overview of grammatical and lexical aspect. A concise review of the pertinent literature is given in section 3. The following section 4 describes the development of the English Verb Analyzer. As this project is on-going, the conclusion 5 focuses on future work.

2 Background

2.1 Grammatical tenses

Finite verbs can carry tense. In the grammatical tense system of English future, progressive and perfect forms are not considered as tense, but as aspect. In this sense, tenses are objective while aspect is subjective and changes according to stance adopted. Prototypical future forms are created and taught using the modal auxiliary verb *will*, but there are many other ways to refer to the future.

Textbooks designed to help non-native speakers of English make extensive use of twelve verb forms or grammatical tenses (Yule, 1998, p.54) shown in Table 1. Grammatical tenses are the twelve permutations of three time periods (present, past, fu-

Tense	Realization
Past	
simple	did
progressive	w(as|were) doing
perfect simple	had done
perfect progressive	had been doing
Present	
simple	do
progressive	(am|are|is) doing
perfect simple	ha(s|ve) done
perfect progressive	ha(s|e) been doing
Future	
simple	will do
progressive	will be doing
perfect simple	will have done
perfect progressive	will have been doing

Table 1: Twelve grammatical tenses for the verb *do* in active voice

ture) and two aspects, namely progressive and perfect (Quirk and Greenbaum, 1993). Eight of these forms are tensed (past or present) while four are modalized (future).

2.2 Grammatical and lexical aspect

Verkuyl (1989) investigated aspectual classes and noted the intricate relationship between semantic information conveyed by aspectual features and verb types, which were defined by the status of the state, i.e. state of no change (state verbs), state of change (process verbs) or change of state (event verbs). Moens and Steedman (1988) showed the complexity of linking temporal expressions to points or periods in linear time as depicted by timelines. Both classifications use different terminology, but align closely with modern descriptions of grammar as detailed below.

Descriptive grammars (e.g. Carter and McCarthy, 2006 and Quirk and Greenbaum, 1993) identify two types of grammatical aspect, perfect and progressive (or continuous). However, aspect is not limited to grammar, words themselves have aspect, specifically lexical aspect. Each verb behaves differently. To distinguish between the different behaviours the verbs are classified into two main groups or classes: stative or dynamic. This aspect affects the meaning of the verb. It is akin to assigning a class to an object in object-oriented programming, and then assigning attributes. In the same way as computer code, subclasses inherit attributes. There are subclasses of both stative and dynamic verbs. Each of these subclasses of verbs sometimes behave the same as their class and sometimes behave in the specific way of their subclass. One of these "behaviours" is realised in the combination of the grammatical tense and the class, explaining its grammatical meaning in context. Verb class affects both syntax and semantics of grammatical tenses. Dynamic verbs are subdivided into durative or punctual. Durative verbs are further subdivided into activities or processes. Each verb class behaves differently. The example below illustrates two different meanings of *present perfect progressive*.

1. I have been walking. (durative verb with focus on process)
2. I have been jumping. (punctual verb with focus on repetition of action)

Therefore, learners of English need to understand how verb class and tense interact to realize grammatical meaning in context.

3 Related works

The extant literature on the identification of tenses for pedagogic purposes is sparse. Even in the broader areas of computational linguistics, tense identification has attracted little attention.

Due to the Zipfian distribution of tenses (Biber et al., 1999, pp.456–463) software that assigns simple present to each finite verb will achieve approximately 55% to 60% accuracy depending on the genre of the text. Software that could accurately predict the most commonly used three tenses and fail to identify any of the remaining nine tenses, could also reach accuracy levels around 90%.

Syntparse claims to annotate tenses with a high (but unstated) degree of accuracy, but this software is no longer available (Lumsden, 1994; Georgiev, 2005).

TimeML (Pustejovsky et al., 2003) is the most well-known system for annotating events and temporal expressions. It is designed for information retrieval purposes, and does not annotate all twelve grammatical tenses automatically.

No online grammatical tense and meaning identification tools were discovered in the research literature.

In earlier work, a tense and aspect identifier was created with an accuracy of 88% for simple declarative sentences (Blake, In press(a)).

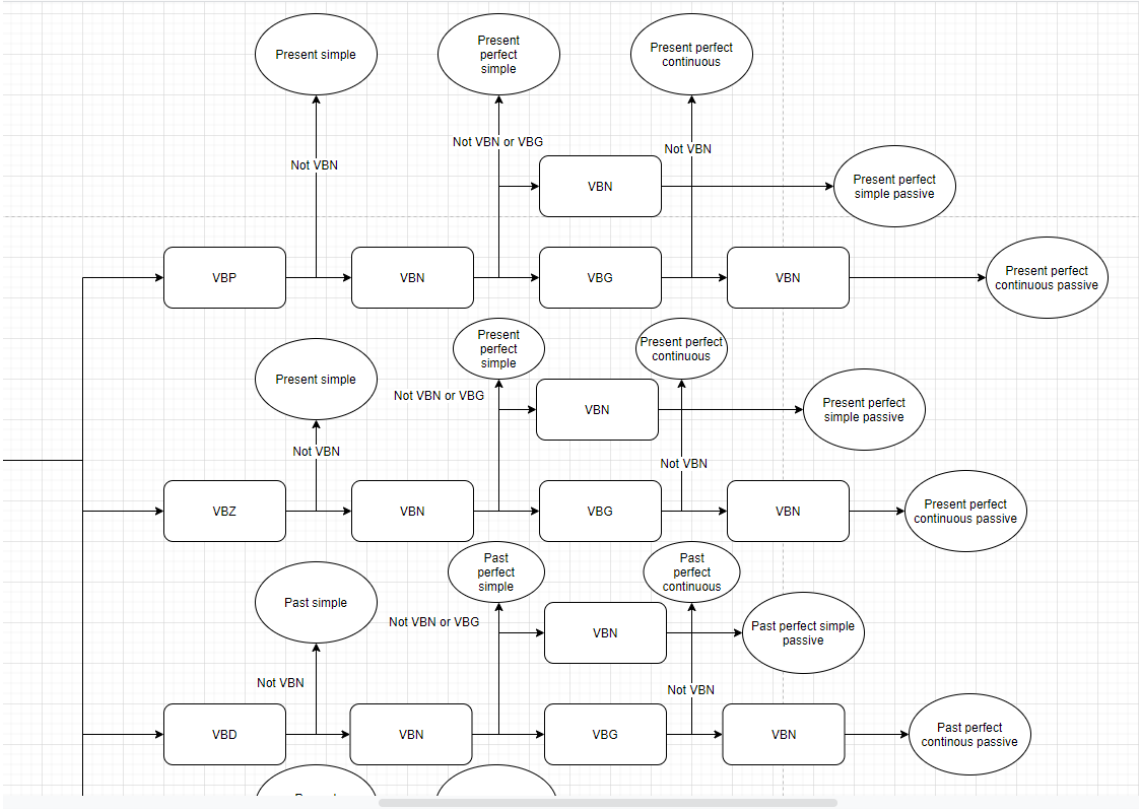

Figure 1: Extract of parse tree diagram using Penn Treebank tagset

Building on this tense and aspect identifier, a tense and meaning identifier was created using a tailormade Python dictionary to store verb classes and subclasses (Blake, In press(b)). This dictionary incorporates verbs from various lists of verb classes and subclasses (e.g. dynamic, durative), some of which are available on github.com/john6938/nlp. A function that matches classes and tense to identify the most frequent grammatical meaning in context was created. However, there were two main limitations to this system. The first was the lack of verb coverage in the dictionary since there is no readily available comprehensive list of verbs categorized by class and subclass. The other limitation is the inability to identify grammatical tenses when auxiliary verbs are elided.

4 Development

The limitations in the earlier works are addressed to a certain extent by using parse trees to match the syntax of tenses rather than conditionals and regular expressions. In addition, in contrast to previous work, this prototype makes use of the supersense function for verbs within WordNet (Miller, 1995) to improve the accuracy of the identification of grammatical meaning in context. Wordnet maps 11,306 verbs to 13,508 word senses (synsets). There are fifteen supersense categories. There is a close, but not exact relationship between some supersense categories and verb types. This version attempts to identify grammatical meaning in context using the verb sense and grammatical tense.

A program was created in Python using the tokenisation and part-of-speech (POS) tagging functions of the Natural Language Toolkit (NLTK) (Bird and Loper, 2004). The tagset selected was the default Penn Treebank tag set (Marcus et al., 1993). A parse tree that specifies the permutations of POS tags that result in particular tenses was created, an extract of which is shown in Figure 1. This parse tree was used as the basis to create a decision tree algorithm in Python. Decision trees start with a node that represents a feature or attribute, which in this case is a part of speech, the branch represents a decision rule (e.g. the presence or absence of a POS tag), and each leaf node represents the outcome (e.g. the tense identified).

A matrix of over 900 permutations of grammatical tense, verb sense and meaning in context was created based on corpus grammars (Biber et al., 1999; Carter and McCarthy, 2006) and confirmed using the web as corpus. Instances were identified

Figure 2: Web app output with tenses colorized

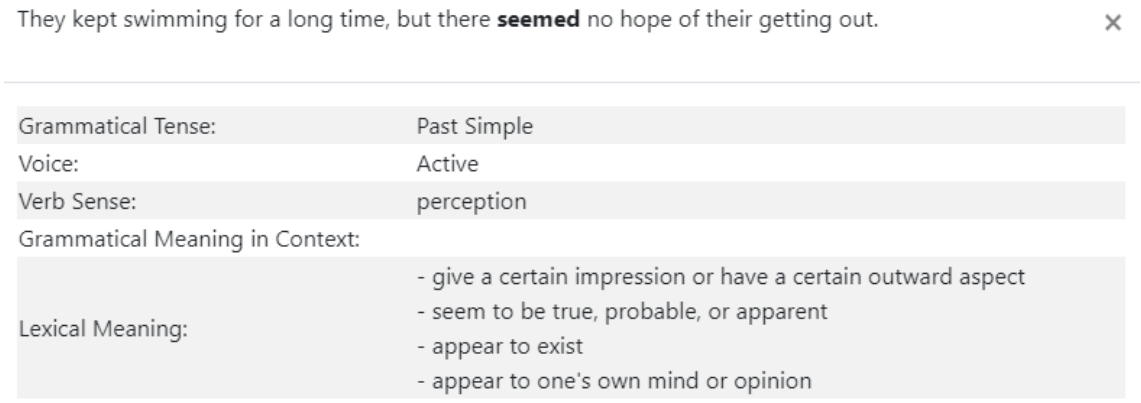

Figure 3: Pop-up explanatory window

for all possible permutations. This matrix provides the verification for the selection of grammatical meanings to display for the 180 tense-sense combinations. A separate Python dictionary was created to store the keys (tense-sense combinations) and values (potential grammatical meanings in context) based on the results of the verified tense-sense matrix.

A graphical user interface was created using the Flask framework. The web app is deployed on the Heroku platform. Users input any text into the user interface. The system highlights and colorizes finite verb groups by grammatical tense as shown in Figure 2. By clicking on a colorized verb group, more information is displayed in a pop-up window (see Figure 3). Currently, this includes voice, grammatical meaning in context, verb sense and lexical meaning. Video explanations will be embedded in this pop-up to help learners understand the tense usage.

5 Conclusion

The identification of grammatical meaning in context is a non-trivial task due to the multiplicity of interrelated issues that can affect meaning. The false positive results in the current tense identification function are caused by inaccuracies in POS tagging. The tense of verbs whose auxiliary verbs are elided, but are not interrupted by grammatical objects can also be identified.

Values still need to be added to complete the key-value hash table that relates verb senses to potential grammatical meanings in context.

Acknowledgments

I would like to thank Mr Tran Vu Duc and Mr Simon Pavlik for their invaluable contributions to the codebase. This research was supported by a JSPS Kakenhi Grant-in-aid for Scientific Research (C) entitled "Feature visualizer and detector for scientific texts", Grant Number 19K00850.

References

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English.* Longman.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

John Blake. In press(a). Tense and aspect identifier. In *Proceedings of EuroCALL 2020*.

John Blake. In press(b). Tense and meaning identifier (version 1). In *Proceedings of the International Conference on Computers in Education 2020*.

Ronald Carter and Michael McCarthy. 2006. *Cambridge grammar of English: a comprehensive guide; spoken and written English grammar and usage*. Cambridge University Press, Cambridge.

Hristo Georgiev. 2005. *English algorithmic grammar*. A&C Black.

Michael Darren. Lumsden. 1994. *Syntparse: A program for parsing English texts: By Hristo GeorgiveGood*.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Marc Moens and Mark Steedman. 1988. Temporal Ontology and Temporal Reference. *Computational Linguistics*, 14(2):15–28.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

Randolph Quirk and Sidney Greenbaum. 1993. *A university grammar of English*. Longman.

Richard Schmidt. 2012. Attention, awareness, and individual differences in language learning. *Perspectives on individual characteristics and foreign language education*, 6:27.

Henk Verkuyl. 1989. Aspectual classes and aspectual composition. *Linguistics and Philosophy*, 12:39–94.

George Yule. 1998. *Explaining English Grammar: A Guide to Explaining Grammar for Teachers of English as a Second Or Foreign Language.* Oxford University Press.