Ontology-Based Approach to Academic Style Marker Identification

Viacheslav Lanin and Sofia Philipson

National Research University Higher School of Economics, Perm, Russian Federation
vlanin@hse.ru, lyubov.filipson@inbox.ru

Abstract. The article describes the ontology-based approach to systematization and search of academic English style markers. The designed ontology is divided into two levels: the first level provides the information about linguistic terms and the second consists of style markers, which were derived by experts in linguistic. It is suggested to generate lexical-semantic template based on the ontology to identify the list of markers in the text with the help of Domain Specific Language (DSL) technology. Currently, there is JAPE-template (Java Annotation Patterns Engine) of GATE text processing system.

Keywords: Style marker, Scientific paper, Ontology, DSL.

1 Introduction

The contribution of research results in scientific publications is the most significant performance indicator of scholars and research co-workers. Papers written on English language notable extend the audience but the scholars, who are not native speakers, usually face some difficulties connected with strict style and language requirements of written academic English. There is huge variety of methodological materials on written academic English as well as specialized educational courses. Literature analysis has shown that suggested recommendations are not systematized and sometimes even have obvious internal contradictions. It should be appreciated that many publications have its own stylistic "publicistic traditions", which are needed to be taken into account while preparing materials. At the moment text investigations are undertaken with the use of computer technology. This enables the processing of huge corpus. Corpus data gives empiric material which can be the foundation for the creation of etalon language patterns, the study of language consistency, and the description of linguistics phenomenon typical of a particular language area, i.e. derivation of style markers. The statistics, collected from corpus annotating in accordance with the style markers, gives the information about academic English criteria frequency of occurrence and their role in style estimation. These will help to define the style quality level of paper and then form development recommendations. Style markers in this paper are considered as main features of academic English in its linguistics meaning.

The main purpose of this project [1] is the extraction of style markers and interrelations between them, and the designing of the academic English style etalon model.

Investigating of hierarchical relations between style elements are also crucial as it helps to determine their frequency occurrence in English scientific texts and describe usage pattern of these elements on the texts pieces of different levels.

2 Existing Solutions

One of the actively developing branches of theoretical and applied stylistics is a complex analysis of English written scientific papers conducted through the large text corpus of particular science processing and comparative stylistics study carrying out. The comparative analysis of English academic style text quality of author, for whom English language is foreign, offers the greatest challenge of corpus linguistics research and the field of software development for corpus analysis. It is worth to say that the major of English written speech research is performed by native scholars and has declarative character or is based on limited data scope. This becomes a problem because of inability to describe English language of the particular subject area with certainty, to derive key features and to study usage pattern. The usage of computer technologies highly simplifies statistical processing of corpus in linguistic research. System-based quantitative research of written scientific speech with the use of software makes possible the statistical processing of large scientific corpus of almost every domain as well as finding of the existent consistency and identification and systematization of main scientific speech attributes.

At this moment there are a great number of tools for corpus processing. The most widespread of them are AntConc, WordSmith Tools, Gate Developer, Sketch Engine and CQPweb. There are specialized solutions for academic papers style analysis, for example project Fapas (Full Automatic Paper Analysis System) [2].

It is also possible to find projects connected with the creation of ontologies, which describe linguistic domain. One of them is GOLD ontology which is General Ontology for Linguistic Description [5]. It gives the description of linguistic basis including most foundational categories and relation between them. The ontology is connected with SUMO ((Standard Upper Merged Ontology) is based on four main domains: expressions, grammar, data constructs, and metaconcepts.

The category expressions mean the physically accessible aspects of language. The base for this aspect was taken from SUMO and to the concept LinguisticExpression were added new ones like WrittenLinguisticExpressions and SpokenLinguisticExpressions.

Grammar category includes the abstract properties and relations of language, the domain that is of primary interest to linguists. It means that anything expressed by a grammatical system be represented by the concept GrammaticalCategory.

Data constructs are constructs that are used by linguists to analyze language data, such as paradigms, lexicons and feature structures. Metaconcepts are the most basic concepts of linguistic analysis, including language itself. There are many ways in which language can be viewed and without a working concept of language, an ontology cannot be used to describe and compare data from all of the world's languages. Language was defined as the set of data associated with a common grammatical pat-

tern. All in all, the ontology tries to describe all the aspect of the natural language which can be applied to all languages.

Another example of ontology is also from linguistic field but it is concentrated on computational linguistics. Developed ontology is built on the basis of scholarly knowledge ontology and because of it concepts of ontology is divided into five hierarchies "whole-part" which are connected to each other with associative relations. Subject of investigation of computational linguistics are the properties and the systems of linguistics units, operations, connected with their functioning in the process of communication, and application processes replied to defined request.

3 Theoretical approach background

The theoretical foundation of the system described in this paper consists of a list of style markers that were selected from reference and study materials, Internet resources about academic writing as well as scientific papers on this topic. All markers from this list can be divided into three main groups: lexical markers, grammar markers, syntactic markers.

Lexical markers include three types of features:

- specific words and terminology (high frequency of terminology; usage of abstract semantic verbs, desemantisized verbs, intensifying adverbs; low frequency of personal pronouns you, he, she, etc.);
- words corresponding to specific word-formation constructions (nouns with -or suffix, commonly used in terminology; abstract nouns derived by suffixes -ment, -ness, -tion, etc.);
- words of specific part of speech (high frequency of nouns, low frequency of pronouns).

Two types of features that fall into grammar markers category are:

- wide usage of verbs in Passive Voice;
- presumable prevalence of verbs in Present Tense.

Syntactic markers can also be classified into two types:

- features described by syntactic structures (simple, complex or compound sentence structure; prepositive and postpositive attributes by most of the nouns; possible prevalence of prepositive attributes in technical texts);
- specific conjunctions, linking expressions, etc. (subordinating and correlative conjunctions; archaisms thereby, therewith, hereby; prepositional phrases; means of logical expressions).

Most of these features can be automatically annotated using lexical-syntactic patterns, although absolute accuracy cannot be guaranteed, which is why expert control and means of manual annotation correction is highly desirable for the system implementation. Flexibility of the system components is also important for development and further testing and debugging due to specificity of academic style feature tagging and natural language processing in general.

Currently our system annotates text based on all of the described style markers with the exception of terminology and sentence structure. Although some components

are still being tested, recent resulting annotation sets provide enough information to analyze academic writing and deepen the studies about some of the features.

For the present style markers are represented as desperate data set. There emerged a necessity of markers systematization besides the method of systematization should give the opportunity of enlargement and adaptation, as language is dynamic and always developing system.

4 Academic style marker ontology

In the present study, a way of regulation and systematization of disparate data set called ontology is going to be described. The ontology reveals the dependences between entities in the form of style markers, and if there are any interconnections, they are indicated. Thus, a huge variety of different style markers turn into a controlled system which then can be used as a part of larger project focused on improving the quality of text annotating.

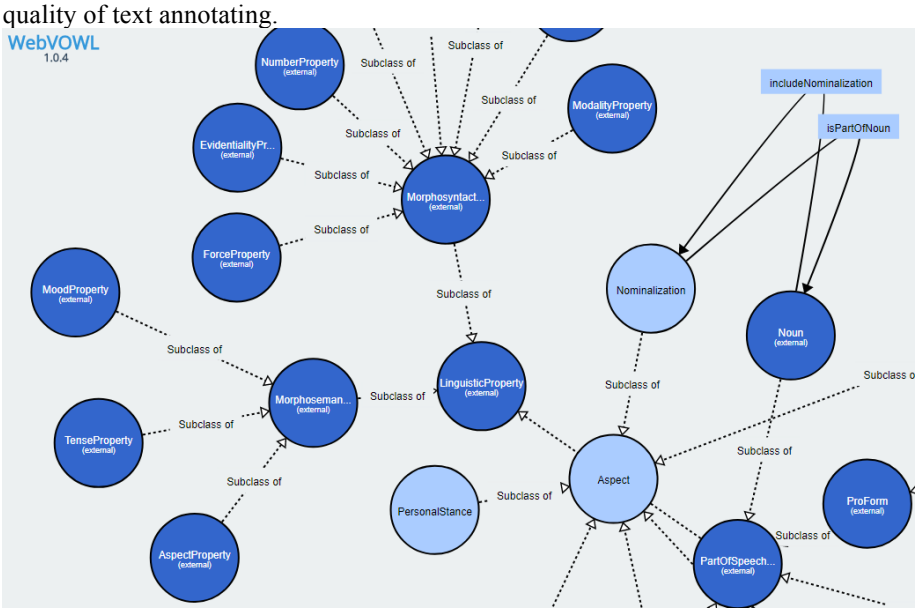

Fig. 1. Visualization fragment of ontology concepts

The ontology is based on the main definitions or basic aspects of academic English which were derived by experts. They are *Nominalization, Personal Stance, Verb, Adverb, Attributes,* and *Cohesiveness*. While adding new classes there was achieved class hierarchy consists of 37 classes and subclasses. The ontology as has been already said has two levels: the level of linguistic terms, which includes such classes as PartOfSpeech, PartOfWord, GrammarStructure, Atributes, and the level of style markers concepts like ComplexConjuctors, PrepositiveAttributes, DesemanticisedVerbs etc.

There are different properties for identifying relations between entities. The main relation is inheritance, which is used for generalization and specification, but also there are properties like hasSuffix (it is the relation between classes Noun and Suffix), depend/influence (between verb and nouns/pronouns) etc.

All in all, developed ontology collects all the derived style markers and reveals relations between them what makes the process of working with style markers simpler.

5 Pattern generation on the basis of DSL-technologies

Ontology is needed not only for style markers systematization but also as the foundation of lexical-semantic patterns generation. Rule generation architecture is demonstrated on Fig. 2.

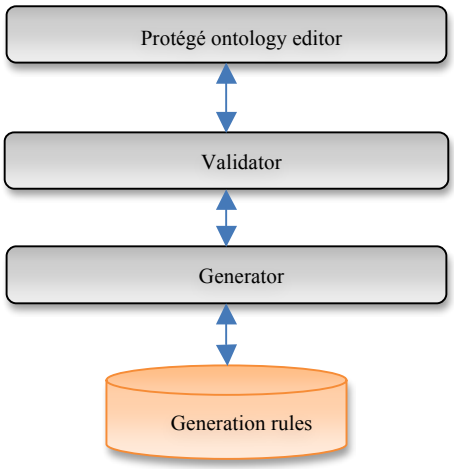

Fig. 2. The architecture of the lexical-semantic pattern generator.

Protégé ontology editor is used for ontology describing and its representation in the OWL format. Validator is the component which is meant for accuracy check of user's models. While designing a model, the user can make some mistakes or make models which are not satisfy the ontology limits constraints. Generator is the component responsible for code generation on target language. Generator is used for transformation of user's models into textual representation on the description language of lexical-semantic patterns as well as file generation into the formats of the computer linguistic systems for example JAPE. To extend the interoperability ability the system gives users the opportunity of determining the transformation rules by themselves. It is crucial on this level of metamodel to make text pattern for every language elements in accordance to which code generation would be implemented. Text pattern includes

the statistic part which is not depend on certain model and the dynamic part, which makes possible the reference to attributes values of different DSL-constructions..

6 Conclusion

Current version of designed ontology consists of 37 concepts and 8 types of relations. The standard tools and software applications are used while designing the ontology which simplifies the process of development and decision maintenance process. The described approached has an expanding property i.e. in order to add new marker the user need to add its description and the identification rule will be generated automatically. Moreover, the use of this linguistics level, which is described in ontology, makes possible the description of related domains.

Acknowledgment

The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2017(grant № 17-05-0020) and by the Russian Academic Excellence Project "5-100".

References

1. Borovikova O.I., Zagorul'ko Yu.A., Zagorul'ko G.B., Kononenko I.S., Sokolova E.G. Razrabotka portala znanij po komp'yuternoj lingvistike // Trudy 11 nacionalnoj konferencii po iskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2008. – M.: LENAND, 2008. –V.3. –p.380-388.
2. Strinyuk S. A., Shuchalova Y., Lanin V. Academic Papers Evaluation Software, in: Application of Information and Communication Technologies (AICT), 2015 9th International Conference on, 14-16 Oct. 2015. Rostov-on-Don: IEEE, 2015. p. 506-510.
3. Scholz T., Conrad S. Style Analysis of Academic Writing // Natural Language Processing and Information Systems: 16th International Conference on Applications of Natural
4. Cunningham H., Maynard D., Bontcheva K., et al. Developing Language Processing Components with GATE Version 7. The University of Sheffield.
5. Farrar S., Langendoen D. A linguistic ontology for the Semantic Web. 2003. GLOT International. 7 (3), p.97-100.