Processing Specialized Terminology in Multilingual Applications: An Interactive Approach

Christina Valavani, Christina Alexandris(✉), Stefanos Tassis, and Antonios Iliakis

National University of Athens, Athens, Greece
cvalavani@hotmail.com, calexandris@gs.uoa.gr, {stassel,antiliak}@di.uoa.gr

Abstract. A Controlled-Language like approach with the integration of expert knowledge is applied to the pre-editing or post-editing of specialized terminology from international texts processed by the UNL System developed by Institute of Advanced Studies of United Nations University (UNU) in Tokyo, Japan. We provide an all-purpose interactive framework focusing on the automatic analysis, ambiguity resolution and editing of German financial terms and English military terms in respect to the Greek language.

Keywords: Sublanguages · Multiword terms · Universal Words · "Safety Mode" Interactive Analysis

1 Introduction

The present approach concerns the design of an interactive editor for managing terminology processed by multilingual Machine Translation Systems, Wordnets or other types of multilingual Natural Language Processing Systems, including the Universal Natural Language (UNL) System concerning the use of Universal Words [7, 8, 11]. A Controlled-Language like approach with the integration of expert knowledge is applied to the pre-editing or post-editing of terminology from international texts, in this case, texts processed by the UNL System developed by the Institute of Advanced Studies of United Nations University (UNU) in Tokyo, Japan. Here, we provide an all-purpose interactive framework focusing on the automatic analysis, ambiguity resolution and editing of two different cases of specialized terminology and language pairs, in particular, German financial terms and English military terms in respect to the Greek language.

Most types of problems encountered and classified concern the morpho-syntactic and the lexical-semantic level of linguistic analysis. Problems related to the latter level observed are mostly affiliated to inherent ambiguities in the sublanguages of financial terms and military terms, often related to different domains such as politics, Information Technology (IT), the law and the natural sciences. Furthermore, since the UNL System processes languages as diverse as Chinese, Japanese, Hindi, Russian, Portuguese and English, un-aided automatic ambiguity resolution based on morphological elements such as case or prepositions may be problematic.

C. Stephanidis (Ed.): HCII 2015 Posters, Part I, CCIS 528, pp. 662–668, 2015.
DOI: 10.1007/978-3-319-21380-4_112

The proposed editing and ambiguity resolution concerns the (1) interactive morphological analysis, especially of compound words and multiword terms, as well as (2) the signalization of predicted sublanguage types and possible differences in semantic content presented to the User. It should additionally be noted that the tool provides the option of registering the statistics of Users choices. The registered choices of the User are assigned a respective probability, allowing the tool to adapt to the type of specialized texts processed.

2 Interface Design

Apart from highly specialized contexts, financial terms and military terms occur in texts intended to non-specialist professionals and/or decision-makers, such as journalists, politicians and managers. These texts often contain terminology from more than one specialized domain, often resulting to ambiguities and other complications in translation or processing due to an overlap in the related fields of specialized terms. The degree of complexity in processing such text types increases when multilingual texts and an international public are involved [1]. In this case processing must be characterized by the possibility to manage terminology from multiple domains (1) with precision and correctness (2), speed and efficiency (3).

To meet the above-presented needs, the proposed design is characterized by three features, namely (1) processing in blocks, allowing (2) a language-independent and Controlled –Language like presentation of analysis, ambiguity resolution and identification of terms with implied or "hidden" components, (3) with a "fast-tack" option and a "safety-mode" option.

The first feature is the analysis of the incoming text in separate blocks, corresponding to the existing paragraphs, since in these types of texts (financial or military domain) a different type of information and sublanguage may be predominant in different paragraphs. The type of sublanguage is identified in each block, allowing the identification of multiple sublanguages and their separate treatment and processing. Text is analyzed in blocks for identifying the sublanguage which is, subsequently, linked to the User's choice to activate the option of "fast-track" or "safety-mode", namely a partially interactive or a fully interactive mode.

The "fast-track" or "safety-mode" option constitutes the second feature of the proposed design. The "fast-track" option is a partially interactive mode to be chosen by Users with expert knowledge of the predominant sublanguage type identified and/or whose native language shares many common features with the language of the text processed, for example, English and Dutch. The "safety-mode" option is a fully interactive mode to be chosen by Users with a general knowledge but no expertise in the predominant sublanguage type concerned and/or whose native language shares few common features with the language of the text concerned, for example Greek, and Chinese. The "fast-track" option can be activated according to User's choice, allowing the adaption of the present interface to various levels of expertise.

As a first step in the interaction, the User inserts the text in the respective field of the designed editor. The module signalizes the sublanguage type with the largest percentage of terms, indicating candidate sublanguage types.

In the second step of the interaction, the User chooses the sublanguage type and then either activates automatic processes related to the chosen sublanguage or selects any problematic words constituting specialized terminology. In this case, the module activates a "stepping stone" or "safety net" mode, similar to strategies employed in Speech User Interfaces (SUIs) [4].

Step in Interaction	**Function**	
"Insert text"	Options:	
"Choose sublanguage type" [Show Sublanguage tags] (SUBLANG-TYPE)	SAFETY-MODE Interactive Analysis (SMI-ANALYSIS)	FAST-TRACK Automatic
"Present statistics" (optional)		

Fig. 1. "Fast-track" or "Safety-mode" Options in Interaction

In the "stepping stone" or "safety net" mode, problematic words are signalized according to ambiguity type, either in respect to a sublanguage or concerning the semantic and morphological analysis, especially in the case of compound words or multiword terms ("Safety Mode" Interactive Analysis –SMI Analysis). In the latter case, the User choses the appropriate type of analysis presenting relations between words and respective readings.

The chosen relation between words and respective reading by the "Safety Mode" Interactive Analysis (SMI Analysis) may be converted in a phrase or sentence that can be easily processed by the Universal Natural Language (UNL) System or any multilingual Machine Translation or Natural Language Processing System, enabling an easier processing of different languages and language families. The User can exit the "stepping stone" or "safety net" mode and return to the automatic processing of the sublanguage and terminology.

The presentation of the statistics of the Users choices registered by the module, assisting the interactive process for future use, is an optional step in the interaction. Thus, the designed editor may be adapted to User behavior and needs and customized to User requirements [3, 9] (Fig. 1).

3 Signalizing Sublanguage Type

The identification of sublanguage type in each block (Sublang-Type) allows the separate treatment and processing of multiple sublanguages, as well as sublanguage-related error prediction and ambiguity resolution. The language-independent SMI Analysis, based on

analytical forms employed in Controlled –Languages, includes the identification of "hidden" compound or multiword terms with the signalization of the chosen sublanguage. "Hidden" compound or multiword terms can be subjected to interactive analytical processing, if requested by the User. Thus, ambiguities may be resolved, both across sublanguages and within the same sublanguage. For selected sets of words, including multiword terms, a list of candidate sublanguages is signalized with the respective tag. This is of special importance in specialized terminology appearing in texts such as political and journalistic texts. The User decides the type of sublanguage activated, if multiple sublanguages are concerned.

In particular, the signalization of the sublanguage type helps resolve ambiguities, for instance, in the case of terms such as "unit", which corresponds to different concepts in the sublanguages of Finance-Economy, Computer Science – IT and the Military. Different domains corresponding to different sublanguages are already indicated in the structure of the Universal Words within the UNL framework [7, 8], as well as in systems such as OWL [5] or in other applications and language resources. Especially in languages highly productive in compound words, such as German, sublanguage type signalization helps identify and separate fixed expressions and specialist terminology from other types of compound words and multiword expressions (such as "ad hoc" compounds in German [2]), the latter often requiring a morphological analysis for the convenience of international Users.

Furthermore, the signalization of the sublanguage type helps identify the identification of "hidden" compound/ multiword terms. For example, in the case of the multiword term "Unions-Wirtschaftsfluegel" in German (retrieved from financial texts), the implied or "hidden" component is the word "Partei" – "political party", the word "Union" corresponding to different semantics according to sublanguage and context. The multiword term "Unions-Wirtschaftsfluegel" (literally "business wing of the Union party") requires expert knowledge for further processing and/or translation in the target language. Another example is the military term "patrol", where the "hidden" component is the word "unit", as a part of the compound term "patrol unit". In addition, the term "patrol (unit)" is differentiated from the verb "(to) patrol [10].

The identification of a chosen sublanguage type allows the automatic processing of a specific sublanguage to be activated, if applicable. If a sublanguage is chosen and activated, the module may function as a "stepping stone" or "safety net" [4], if problematic cases are encountered. If ambiguities occur within the same sublanguage, differences in semantic content and possible readings are presented to the User. This is the most usual case with compound words and multiword terms.

4 Morphological Analysis and the UNL Framework

Morphological analysis of compound words or multiword terms, including "hidden" compound or multiword terms (SMI Analysis) takes place in the "safety-mode" option of the interaction. The proposed editor is activated with the identification of words constituting specialized terminology identified by the User.

In order to ensure a simplified form of an independent analysis across languages and language families, the presented analyses are neither identical nor similar to but compatible with the presentations of logical relations with the Universal Word –UNL framework [7, 8, 11]. Compound words or multiword terms are analyzed according to the parameters of the natural language concerned. For example, in German compound terms, and multiword terms, the component constituting the "head" of the term [6] is identified and a set of possible relations between the components is shown, with the indication of the most probable reading. This feature is of special importance in cases where neologisms are concerned or in very productive languages such as German, where there is a continuous formation of new ("ad hoc") compound words.

GERMAN-INPUT (TEXT):	SMI-ANALYSIS:	ENG- TERM:
Preisschlacht [^{M}der Preis+ ^{F}die Schlacht-HEAD] GR-ANALYSIS: Μάχη των τιμών [^{F}η μάχη (the battle)+ ^{F}η τιμή (the price)]	[war/battle OF prices] [war BETWEEN /ABOUT prices]	price war
Umsatzrückgang [^{M}der Umsatz$^{\pi}$ + ^{M}der Rückgang-HEAD] GR-ANALYSIS: Πτώση των πωλήσεων [^{F}η πτώση (the fall)+ ^{F}η πώληση (the sale)]	[decline/fall OF sales] [sales SUBJECT decline VERB] [lowered sales]	decline in sales (or:) sales decline

Fig. 2. SMI Analysis of German Financial Terms

The types of analyses depend on the types of natural languages concerned. In the case of German, multiword compound words are identifiable, as most components occur as a single string, in some cases separated by a dash. By applying morphological parameters, the head of German compound words and multiword terminology is identified to the right [6]. We note that English and Greek expressions constituting multiword terms do not occur as a single string and are not always easily identified, even by human editors or translators (Fig. 2).

Alternative analyses are presented in the interface, depending on what components in the multiword terms are chosen to have a closer relation each other. Presented SMI analyses are a simplified form of logical relations between components, for example, "BELONGS-TO", and "OBJECT-OF-ACTION". Furthermore, as in analytical forms of Controlled –Languages [10], multiword terms can be converted into sentences.

The statistics of Users choices for the possible analyses presented are registered and assigned a respective probability, as a stated above (Fig. 3).

UW GENERATED TERM:	SMI-ANALYSIS:	GR - TERM:
[Στρατιωτικών Πληροφοριών] "Military Intelligence"	[Service FOR Military Intelligence] => (Identified by SUBLANG-TYPE)	[Υπηρεσία Πληροφοριών Στρατού] "Military Intelligence"
[Στρατός Assault Team] "Army Assault Team"	[Team FOR Assault OF Army] [Team SUBJECT-OF-ACTION Assault] [Team BELONGS-TO Army]	[Ομάδα Κρούσης Στρατού] "Army Assault Team"
[άμεση μονάδα υποστήριξης] "direct support unit"	[unit FOR direct support] [unit OBJECT-OF-ACTION direct support]	[μονάδα άμεσης υποστήριξης] "direct support unit"

Fig. 3. SMI Analysis of English Military Terms

5 Conclusions and Further Research

Un-aided automatic ambiguity resolution may be problematic for Systems processing diverse languages and language families, especially when remarkable differences in fundamental elements expressing logical relations and grammatical information are concerned. The designed module combines automatic and interactive processing, integrating expert knowledge and Controlled-Language practices, allowing the User to proceed with fast-track editing of the texts with specialized terminology or to activate a "stepping stone" or "safety mode", as applied in SUI Systems. A full implementation allowing a contrastive evaluation in regard to both languages and sublanguages concerned is targeted to contribute to the further development of the approach and interface designed. Further research includes an extension and enrichment of the existing identified tag types and the improvement and fine-tuning of the interactive SMI Analysis.

References

1. Alexandris, C.: Managing Implied Information and Connotative Features in Multilingual Human-Computer Interaction. Nova Science Publishers, Hauppauge (2013)
2. Busch, A., Stenschke, O.: Germanistische Linguistik, Eine Einfuehrung, 2nd edn. Gunter Narr, Tuebingen, Germany (2008)

3. Jurafsky, D., Martin, J.: Speech and Language Processing, an Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, 2nd edn. Prentice Hall Series in Artificial Intelligence. Pearson Education, Upper Saddle River (2008)
4. Lewis, J.R.: Introduction to Practical Speech User Interface Design for Interactive Voice Response Applications, IBM Software Group, USA. Tutorial T09 presented at HCII 2009 San Diego, CA (2009)
5. Loaiza, F., Wartik, S., Thompson, J., Visser, D., Kenschaft, E.: The Best of All Possible Worlds: Applying the Model Driven Architecture Approach to a JC3IEDM OWL Ontology Modeled in UML. In: Proceedings of the 19th International Command and Control Research and Technology Symposium, 19th ICCRTS, Alexandria, VA 16–19 June 2014
6. Sternefeld, W.: Syntax, Eine morphologisch motivierte generative Beschreibung des Deutschen, vol. 1. Stauffenburg, Tuebingen, Germany (2006)
7. Uchida, H., Zhu, M., Della Senta, T.: Universal Networking Language. The UNDL Foundation, Tokyo, Japan (2005)
8. Uchida, H., Zhu, M., Della Senta, T.: The UNL, A Gift for Millennium. The United Nations University, Institute of Advanced Studies UNU/IAS, Tokyo, Japan (1999)
9. Wiegers, K.E.: Software Requirements. Microsoft Press, Redmond, WA (2005)
10. Xue, P., Poteet, S; Kao, A., Mott, D., Braines, D., Giammanco, C., Pham, T.: Information Extraction Using Controlled English to Support Knowledge-Sharing and Decision-Making. In: Proceedings of the 17th International Command and Control Research and Technology Symposium, 17th ICCRTS, Fairfax, VA, USA, 19–21 June 2012
11. The Universal Networking Language http://www.undl.org