Utilizing NLP tools for the creation of school educational games

Aristides Vagelatos[1[0000-0001-7825-0550]], Monica Gavrielidou[1], Maria Fountana[1] and Christos Tsalidis[2]

[1] Computer Technology Institute & Press, Athens, Greece
[2] Neurocom S.A., Athens, Greece

Abstract. The use of digital games to support learning (game-based learning) through an alternative, more attractive way is rapidly growing in both European and worldwide educational sector. Digital games are a rapidly developing field, as they are amongst the most popular technologies that young people use for their entertainment. Within this framework, project "Lexipaignio" was initiated in order to develop an innovative and state-of-the-art NLP (Natural Language Processing) environment for the creation of digital educational games for Greek students. These games will be dynamically generated by the educator for his/her students, in order to improve various vocabulary and linguistic skills, while understanding the context of specific school subject areas. In this paper we present the NLP environment for the subject of Geography.

Keywords: Educational Games, NLP, Game-based Learning.

1 Introduction

Lately, with the integration of new technological achievements into both educational and everyday life of students, important changes are under development in the educational and learning processes, where the Information and Communications Technology (ICT) plays a major role. From this perspective, the use of digital games to support learning (game based learning) in a more attractive way is rapidly emerging. Digital games are a rapidly developing field, as they are amongst the most popular technologies that young people use to amuse themselves [7]. The educational potential of digital games is correlated to motivation, amusement and the trigger of interest, which are considered consistent with positive learning results. According to relevant research, computer games provide a quick and interesting learning pace in contrast to the conventional teaching methods and in this perspective, they can evolve into a challenging way as far as digital learning is concerned [3].

In the "Lexipaignio" project, an innovative and state-of-the-art computational environment is under development, through the creation of digital educational games for students of upper primary and lower secondary education in order to: a) improve their language competency level and overall linguistic abilities, b) develop various

vocabulary and linguistic skills, while deepening in the context of specific school subjects (biology, geography etc.).

At the point of convergence of computer science and linguistics, Natural Language Processing constitutes a challenging area for numerous applications in our everyday life. Focusing on Education, the "Lexipaignio" project aims at the use and further development of a series of Natural Language Processing infrastructure tools (Morphological Lexicon, Lemmatizer, Mnemosyne language editing system, corpus of Greek school subjects, etc.), for the implementation of dynamically created gamified educational material. This paper emphasizes on the utilization of NLP tools for the creation of minigames regarding the subject of Geography in Greek secondary schools. As part of an ongoing project, the development of Geography minigames aims at providing useful feedback regarding the use of Natural Language Processing for the development of dynamic gamified materials in other school subjects.

In the rest of the paper, the developed NLP infrastructure is presented, then, the selected test bed (Geography) is described, and finally the drawn conclusions are being quoted.

2 NLP Infrastructure

Natural Language Processing (NLP) is a research area that gains extreme interest in the post-information age. The ability to process information and transform it to knowledge is considered essential in today's "information jungle" [2]. In the context of education it seems that NLP has great benefits to offer [1]. In this section, the necessary NLP infrastructure is examined in order to support the dynamic creation of educational games within the project's scope. Most of these tools are based on earlier research work which has been further developed to constitute the basis of the "Lexipaignio" project.

2.1 Greek language

In specific, statistical language models that select "word terms" do not work well with Greek text due to its rich morphology, returning poor results for both "precision" and "recall" metrics [9]. For single word terms, we must handle the morphology of words along with the part of speech. With more than 20 word forms for adjectives and more than 200 word forms for verbs, counting frequencies of words does not provide good results. A combination of POS (Part Of Speech) filtering, lemmatization (normalization) and TF/IDF scoring give the best results for single word terms. Equally, multi word terms have similar problems: e.g., a 2-tuple sequence of the form |ADJECTIVE NOUN| can have more than 100 different forms and only a few of them can be considered as valid terms. In this case, lemmatization and POS filtering does not help. For n-tuples with n > 2 the challenge remains. Thus, we use "KANON" formalism [4] to recognize valid patterns of candidate terms, e.g. |ADJECTIVE_GCN NOUN_GCN| where _GCN suffix codifies the G=Gender, C=Case, N=Number, i.e. the adjective and noun must agree in gender, case and number in order to be considered as candidate terms. Recognizing a candidate term is the first step of filtering. The second step is to

normalize the expression by lemmatizing both words. These filtered expressions then feed the n-tuple language model for the statistical evaluation of candidate terms.

Another challenge is the "compound words" which are common in the scientific terminology of the Greek language. Together with morphology and extensive vocabulary they increase the difficulty of classifying the texts and finding common meanings. To handle this phenomenon, we split the words in q-gram characters and take these q-grams as units instead of entire words. For example, the compound words "geology" and "geography" have the word "geo" common and if we want to utilize this fact in a classification process we must use 3-grams ("geo") or 4-grams ("_geo") as units.

2.2 Corpus processing

Within the projects' framework, a geographical corpus was compiled (see next section) to serve as a basis for the rest of the tools as well as input for the educational games, after being annotated. The annotation process of the LEXIPAIGNIO corpus involved almost all NLP components that already existed and included: a tokeniser, a sentence splitter, a morphosyntactic tagger, a toponym gazetteer and a single and multi-word term recognizer.

The collection of documents used includes more than 180 chapters from the educational books used in upper primary and lower secondary education in Greece, cleared from problematic typographical and layout related elements. Then a number of NLP tasks was applied in order to find ways to utilize the content in the games. Specifically:

Candidate terms extraction by applying a set of morphosyntactic patterns. We used 20 morphosyntactic patterns expressed with "KANON" rules that recognise candidate terms. The distribution of terms in these 20 rules along with the frequency for each candidate term as well as the frequency of the prefixes, provide valuable information about a) the most common morphosyntactic patterns of terms b) confidence level and c) patterns for synthetic (fake) terms that can be used in games. Table 1 shows the number of candidate terms per pattern as well as the lengths (number of words) recognized by each pattern.

Table 1. Candidate terms

Pattern	#Words	#Candidate Terms	#Instances
1	2	4845	8272
2	2-8	1200	1484
3	2-12	912	1199
4	3-11	215	239
5	3-8	410	481
6	3	383	433
7	3-7	76	81
8	3	1042	1185
9	4-7	62	62
10	4	79	80

11	4	17	17
12	4-10	56	60
13	4-10	118	118
14	4-8	72	77
15	4	91	99
16	5-7	12	12
17	5-9	14	14
18	5	8	8
19	5	7	7
22	3-10	250	256

Classification of chapters. Two unsupervised classification algorithms were applied in order to discover the thematic categories of the chapters: a) *K-Means* with *Euclidean distance* as metric and b) *Hierarchical clustering* with *cosine similarity* as metric [8]. To minimize the problem of rich morphology and to consider common stems of compound words we used three types of elements: words, 3-gram characters and 4-gram characters. In Table 2 we present the results of the different clustering methods.

Table 2. Clustering method results

Method	**#Clusters**	**Unit**
K-Means (K=10)	6 (4 empty)	Word
Hierarchical	24	Word
K-Means (K=10)	7 (3 empty)	3-gram
Hierarchical	18	3-gram
K-Means (K=10)	5 (5 empty)	4-gram
Hierarchical	18	4-gram

Discovery of keywords per chapter. Keywords are important in the classification and characterization of a document/chapter. In order to select the appropriate words that distinguish a chapter from the rest of the chapters, the TF-IDF [6] (Term Frequency-Inverse Document Frequency) algorithm has been used in words, lemmas, and truncated words. For each word and chapter, we compute its TF-IDF metric which is the weight of the word for the chapter. It is computed by the product of TF * 1/DF where TF is the number of times a word is presented in a chapter and DF is the number of documents in the collection that contain this word. This means that a word increases its weight for the specific chapter in case it appears many times (frequency) but it loses (minimize) its weight if it also appears in many chapters in the collection (common words). Later on, the Okapi-BM25 [11] algorithm was applied, which is an alternative of TF-IDF, with two approaches: In the first approach we used the lemmas of the words for every chapter and for unknown words, the words themselves have been used. In the second case for unknown words we truncated the words removing the suffix in case it matched a common suffix of the Greek morphology. The words with higher scores were extracted and used as keywords for the corresponding chapters.

2.3 Augmenting Morphological Lexicon

The Greek morphological lexicon (~ 90.000 words, ~ 1.200.000 word forms) that has been implemented in a previous project [9, 10] was enriched with the geographical terminology extracted from the corresponding Corpus. The process of collecting corpus-based geography terms was based on the hypothesis that if a word-form found in the corpus is either unknown to the lexicon or has got high TF/IDF score in the corpus, then there are good chances for it to be a geographical word-form.

2.4 Tools

Several tools were utilized in the process of term extraction. Among these, a concordance was particularly helpful to the linguists that had to manually inspect and correct the output of the automatic terms selection tool. As can be seen in Fig. 1, the concordancer, allowed the linguists to inspect the usage of any term within the corpus in order for them to decide whether or not it constitutes a valid geographical term.

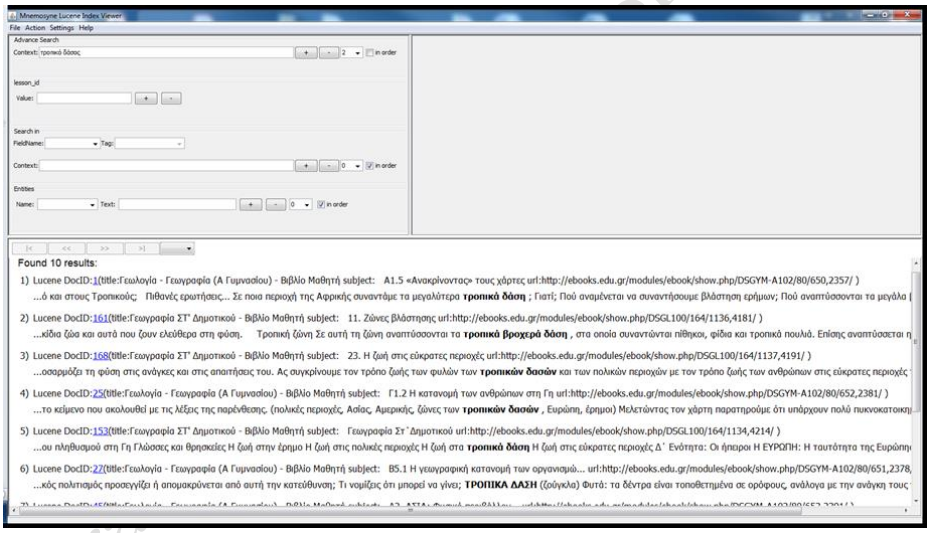

Fig. 1. The concordancer (a screenshot presenting the use of the term "τροπικό δάσος" (tropical forest)).

3 Test Bed: Geography

Among several other candidate school subjects (i.e. Biology, Home Econnomics, Music, Religious Education, etc.), with Modern Greek Language holding the main focus of the project for full deployment of the developed infrastructure, Geography was selected as the subject to be used for the project hypothesis test bed.

Although it was initially estimated that schoolbooks of the first two grades of secondary school would be sufficient to compile a substantial corpus for the subject of Geography, it was decided by the partnership to also explore educational activities from online Greek collections and repositories such as 'Ifigeneia' and 'Photodentro' as well as other relevant educational resources. As a result, the corpus of Geography was additionally compiled by full educational packages of teacher and student books taught both in upper Primary and lower Secondary education in Greece, escorting curricula and complementary educational material which embraced Geology and other related sciences complying with the interdisciplinary structure of the relevant schoolbooks.

Developing the process for the automatic extraction of thematic vocabulary and other structured information from the collected educational material and proceeding with the generation of taxonomies in the field of Geography has been a challenging task. It involved the identification of patterns both in terms of text format and layout, but also of linguistic morphosyntactic structure and resulted in an additional, on the side and beyond the scope of this project, interesting list of suggestions on the structure of future interactive books.

The corpus of Geography was, then, automatically annotated through the use of the NLP tools to automatically extract structured information from the semi-structured electronic documents. Additionally, with the use of the built-in "KANON" formalism, rules were formulated since KANON, as already described (see section 2.2.), is able to define context sensitive constructs that take into account their contextual elements and has the potential to exploit words' individual characteristics (morphological, stylistic, etc.), according to the prior analysis of the language and the recording of its formalistic imprint. The result of this task for the subject of Geography has been the development of a unique resource of monolectic (single word) and polylectic (multi word) terms interconnected to their original sources, upon request, and escorted with additional text related information.

For the "Lexipaignio" project, the corpus of Geography, and primarily the corpus of Modern Greek Language to be developed at a later stage, is a valuable system resource out of which structured linguistic material will be drawn for the development of dynamic, thematic mini games. Consequently, teachers will be able to determine the content of the dynamic mini games in terms of theme and educational level / student age whereas they might also be able to go as far as specifying the relevant book chapters to be used as the primary source for the mini games that they will be releasing to their students.

3.1 Taxonomy of Geographical – Geological Terms

To facilitate the production of a large number of questions suitable for thematic games, a taxonomy of geographical and geological terms was developed based on the thematic vocabulary extracted from the relevant corpus. Expanding the possibilities of the "KANON" formalism by utilizing the syntactic forms which result from the co-occurrence of terms, will allow the automatic creation of classifications and will establish the functionality of formalism for further use in texts and other thematic fields.

For the compilation of the taxonomy, a step towards the formation of an ontology that will constitute the standard description of the geographical – geological knowledge, the following steps were followed:

- Four school textbooks were selected and isolated from the corpus of Geography: two textbooks taught in the senior classes of the Primary School and two textbooks taught in the junior classes of the High School.
- The contents of the above books were merged (overlapping sections among the books were grouped together).
- A first version of the distribution of the material throughout the textbooks was developed and mapped based on the prior grouping and the categorization of their chapters.
- Verification of the functionality of the categorization was attempted to additionally ascertain any overlaps and repetitions.
- Corrections were made to the initial categorisation and the taxonomy was finalized (part of the upper level can be seen in Fig. 2).
- Code numbers were assigned to each category of the classification as well as to each section of all school textbooks.
- The taxonomy was embedded in the project's portal, in the following address: http://lexipaignio.cti.gr/apotelesmata/ypodomes/taksinomia.

The categories of concepts that compile the initial classification concern a wide range of Geography-Geology study areas (e.g. the natural environment, the anthropogenic environment, continents, etc.), which serve the needs of the project effectively. In addition, the range of the categories ensures coherence in the knowledge of the relevant subject area, which can be generalized and reused in a variety of other subfields.

3.2 The concept of mini games

Designing scenarios for educational games is a challenging process that requires creativity, perseverance, tough negotiations, continuous reviews and adjustments.

What has been acknowledged as absolutely critical and innovative in Lexipaignio was not the game per se but the mechanism behind the game that would enable teachers / users to deepen in (i) various grammatical, morphological and vocabulary related phenomena taught at school (ii) by using ready-made, customized or even created from scratch simple and popular mini games, (iii) drawing exemplary sections directly from school text books and (iii) automatically connecting them to specific online tools such as online spellers, grammar checkers and various dictionaries developed and maintained by members of the project partnership.

In the case of geography, the development of the taxonomy of geographical – geological terms, as described above, is seen as the liaison between the mini games and the school text books. The 'game mechanism' under development foresees that the teacher / user will be able to connect sections of the taxonomy to specific, available mini games through a set of predefined, specific rules. Activities approached through the mini games may extend from the exploration of different meanings of specific vocabulary in

geography to the comparison of specific geographical phenomena presented through different school text books addressing students of different age.

Game features such as 'levels', 'scoring', 'time limit' and 'help' aim to turn the mini games more appealing to end users / students. Online 'help', in specific, will constitute an additional way for the mini games to connect to the available, online tools (one simple example of an online help could be to check the meaning of a 'geographical term' on the online dictionary) with the primary connection being made through the set of rules that will be used in order for specific content from the database of the online tools to be 'fed' into a specific mini game (one simple example of such a rule could result in 'feeding' a mini game with all or specific phrases from all geography school text books that contain the word 'water').

3.3 The role of the teacher

Initial shift of the emphasis from one exhausting stand-alone, ready to play, educational game to a collection of multiple, simple, model mini games also changed the perspective of the role of the teacher, who can select to be in charge of game play by specifying the areas or the phenomena (grammatical, morphological, lexical) of focus.

Thus, it is estimated that teachers will be able to either select the suitable mini games for their students from an extended collection that will be available or create a new version of one of the available mini games by specifying the content and/or one (grammatical, morphological, lexical) phenomenon from the school books or the taxonomy, in the case of geography.

Furthermore, mini games will be equipped with extra features in order for teachers to be able to monitor their students' performance and most importantly to be able to provide feedback and assist the learning process whenever deemed necessary.

Finally, teachers will be able to load new material to a mini game (e.g. questions – answers) through ready-made templates whereas the possibility of defining additional grammatical, morphological or lexical phenomena of interest through tagging / annotations is still being explored.

3.4 Intentions of using NLP techniques and pedagogical objectives

In recent years, the presence of NLP technologies in digital games has been eminent. Similarly, NLP is being used in educational digital games. However, there is little research on NLP methodologies in this educational context. Specifically, Picca et al [5] argue that this is related to the fact that NLP techniques are widely used to facilitate the game pedagogical objectives and are not treated as distinct game features. They stress on the importance of clarifying the objectives of using NLP and the pedagogical objectives of an educational digital game. Effectively, these two objective categories should be aligned in order to serve the overall game concept.

NLP technologies form a useful tool for the creation of language-based activities par excellence. In Lexipaignio project NLP technologies are linked to the dynamic creation of a variety of language mini games based on structured linguistic material. Thus, the Geography test bed serves as an example of how NLP technologies can assist in

creating learning tools for the teaching of subjects other than Greek Language. The NLP techniques in the Geography test bed share the same objectives with the development of the Greek language games. At the same time, the pedagogical objectives in the Geography test bed focus on helping teachers combine educational resources on the basis of the geography taxonomy to dynamically create mini games according to their students' needs. Furthermore, the Geography test bed aims at improving students' knowledge on Geography and Geology through the study of terminology and the construction of geographical knowledge (information on the earth natural and human environment) according to the Geography syllabus for the Greek secondary school. The question types used in the geography test bed mini games are multiple choices, gap filling, true/false, put in the right order, sentence synthesis and riddles.

4 Conclusions

With the aim to deploy Natural Language Processing infrastructure for the creation of educational games in a variety of school subjects (Geography, Modern Greek Language, Biology), so far, the language processing techniques applied in "Lexipaignio" project provide encouraging results. With the infrastructure almost completed, the next step is to implement the mini games designed, in order for them to be tested and approved by the educators. Though this is not a straightforward task since it needs a lot of interdisciplinary co-operation as well as extensive "on site" testing, the research so far is more than promising as far as the successful implementation of the project is concerned. This will enable educators to easily create minigames according to their students' needs, by regulating the game content.

Acknowledgment: This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T1EDK-05094).

References

1. Alhawiti, KM.: Natural language processing and its use in education. International Journal of Advanced Computer Science and Applications, 5(12), (2014).
2. Al-Rfou, R., Perozzi, B., Skiena, S.: Polyglot: Distributed word representations for multilingual NLP. In Proc. of CoNLL (2013).
3. Gregory, S., Torsten, R., Wood, L., Henderson, M.: Gamification and digital games-based learning in the classroom. In: Teaching and Digital Technologies: Big Issues and Critical Questions. Henderson, M., Romeo, G. Editors, Cambridge University Press (2015).
4. Kokkinos, Th., Gakis, P., Iordanidou, A., Tsalidis, Ch.: Utilising Grammar Checking Software within the Framework of Differentiated Language Teaching. In Proceedings of the 2020 9th International Conference on Educational and Information Technology, Oxford, United Kingdom, (2020).

5. Picca, D., Jaccard, D., Eberle, G.: Natural Language Processing in Serious Games: A state of the art. International Journal of Serious Games, 2(3), 77-97 (2015).
6. Ramos, J.: Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning, 242, 133-142, (2003).
7. Reinders, H.: Digital Games in Language Learning and Teaching. Palgrave Macmillan Publishing (2012).
8. Singla, A., Karambir, M.: Comparative analysis & evaluation of euclidean distance function and manhattan distance function using k-means algorithm. International Journal, 2(7), (2012).
9. Tsalidis, C., Vagelatos, A., Orphanos, G.: An electronic dictionary as a basis for NLP tools: The Greek case. In Proc. Of 11th Conference on Natural Language Processing, Fez, Morocco (2004).
10. Vagelatos, A., Mantzari, E., Pantazara, M., Tsalidis, Ch., Kalamara, C.: Developing tools and resources for the biomedical domain of the Greek language. Health Informatics Journal, 17(2), 127-139 (2011).
11. Whissell, J.S., Clarke, C.L.: Improving document clustering using Okapi BM25 feature weighting. Information retrieval, 14(5), 466-487, (2011).

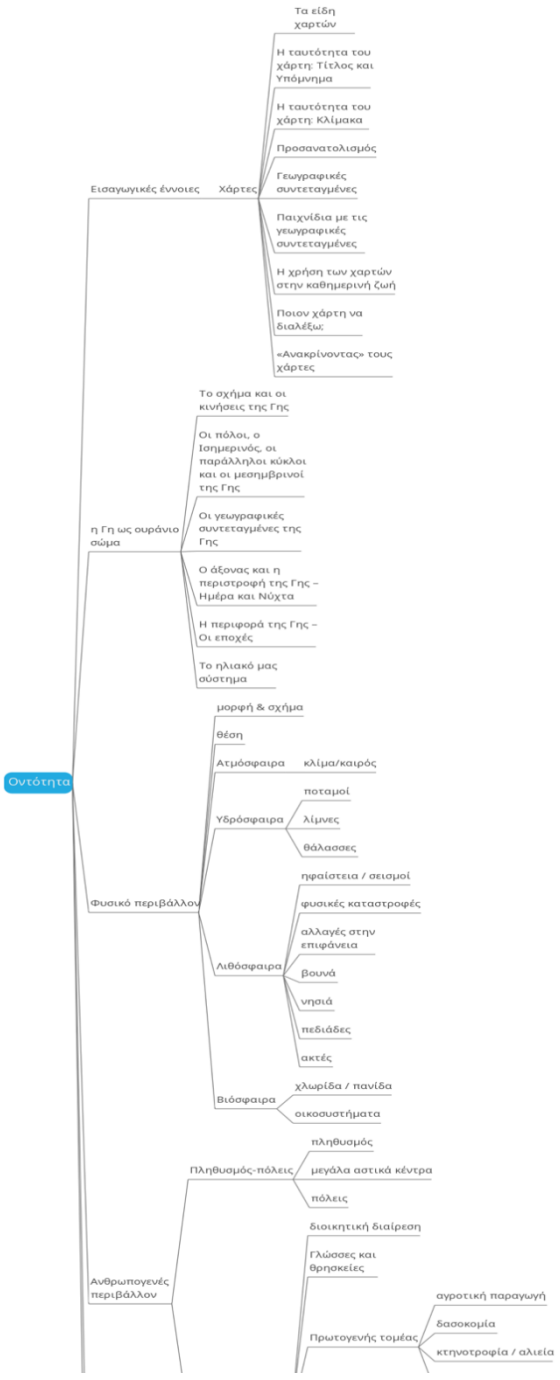

Fig. 2. Part of the upper levels of the taxonomy of Geographical – Geological Terms.