University of South Wales
Prifysgol
De Cymru

AI Marking Tool

Using Natural Language Processing

Michal Stec

Abstract

In recent years, a large amount of studies suggest that teachers are highly overworked. This contributes to lower satisfaction with their job and hinders the quality of their teachings. One of the main areas where the workload could be lowered is the amount of marking that teachers have to go through in a week.

This project explored various tools available to the teachers that could help with automating and speeding up the marking process. Moreover, a semi-automated system had been developed, utilising various python NLP technologies and libraries such as NLTK and WordNet that aimed to achieve the above.

The main focus of the system was to assess the semantic similarity between a short essay provided by the student and a set of criteria produced by the teacher.The software aimed to provide the teacher with a facility that helped produce the criteria and also view and supervise the results produced by the software when measuring the semantic similarity. The main problems that the software had to overcome was to apply Word Sense Disambiguation as well as calculating the Semantic Similarity between a word and its synonyms.

In conclusion, a successful prototype had been developed, that showed promising results in being able to grade a short essay. Calculating the semantic similarity between words and ultimately sentences was achieved with a high success and accuracy. On the other hand however, the current state of the Word Sense Disambiguation algorithms only allowed for around 40% accuracy, making the software dependent on direct word comparison rather than syntactic comparisons which significantly lowers the correctness of the results.

Introduction and Background

In recent years, studies have shown a rising job dissatisfaction amongst teachers. Droognebroeck, Spruyt and Vanroelen (2014, p. 99) argues that a heavy workload, especially non-teaching related (paperwork, assessing work) is one of the main predictors of job dissatisfaction amongst teachers. Survey carried out by the National Education Union (2018) found that primary school teachers spend an average of 59 hours a week working.

Classroom Teachers						
Priamry		Secondary		Academy		
HR	%	HR	%	HR	%	
10.6	17.8	8.5	15.2	8.2	14.8	Planning lesson, test, assessment
9.7	**16.3**	**9.4**	**16.8**	**8.7**	**15.7**	**Assessing pupil work, reports**
2.4	4.0	1.1	1.9	1.9	3.4	Other non-contact activities

Figure 1. The average amount of hours spent a week by a teacher on certain activities (Great Britain. Department of Education, 2014).

One of the objectives of the project was to look into existing products and solutions that could help speed up the process of assessing pupils' work. The existing solutions could be split into two main categories:

- **Manual Scoring Systems** – easy to implement and use programs that offer simple tools that help automate some of the processes involved with marking essays. This included calculating grades and statistics, storing often used comments and allowing easy lookup of criteria (*EssayTagger, e-Marking).*
- **Automated Essay Scoring (AES)** – systems that introduce full automation into scoring and evaluating written text. These systems use Natural Language Processing tools as well as Machine Learning in order to provide automated grades and feedback to the students. These systems have proven to be extremely accurate (*Criterion, EdX)*. Despite that, they have been criticised by the public for a lack of human interaction in the process as well as the requirements of having large amounts of training data which makes it difficult to implement for smaller institutions (Dikli, 2006, p. 1).

The research conducted helped structure the developed software so that it aimed to fall in between these two categories. The goal was to provide supervised automation that would not require large amounts of data in conjunction with a set useful tools inspired by the add-ons and plug-ins.

Methods

The developed software was split into two parts. First is the Frontend which provides the teacher with a facility to create an examination JSON file with criteria and marks available, as well as view and modify the results provided by the grading software. This has been developed using C#.

The second part is the Backend. This has been developed using Python and its Natural Language Processing libraries such as NLTK and WordNet. This part was responsible for automated grading of the essays.

Algorithms

In order to calculate the semantic similarity between two sentences certain pre-processing steps had to be undertaken. First, each sentence was **tokenized** into words and each word was assigned its **part-of-speech tag**.

Next, **Word Sense Disambiguation (WSD)** was applied to every word. This is a process of identifying what sense of a word was used in a given sentence. This was achieved by the *Michael Lesk Algorithm* that uses WordNet's gloss dictionary.

```
PINE
1. kinds of evergreen tree with needle-shaped leaves
2. waste away through sorrow or illness
```

```
CONE
1. solid body which narrows to a point
2. something of this shape whether solid or hollow
3. fruit of certain evergreen trees
```

Figure 2. Gloss terms for word PINE and CONE (Wikipedia, 2018)

The algorithm looks up the gloss terms of given word and its neighbouring tokens. The gloss terms are then cross-checked, looking for any overlap in words. The senses with the highest overlap are then assigned to a given word (Lesk, 1986).

Lastly, **Semantic Similarity** between each sentence and criteria is calculated by comparing the semantic similarity of each noun/verb in a sentence to each noun/verb in the criteria. This is done using the *Wu Palmer Path Length Similarity algorithm*. The software accesses synsets of each wordsense and based on the taxonomy provided in the WordNet, calculates the path distance of each synonym from their *Least Common Subsumer* (Gole, 2015).

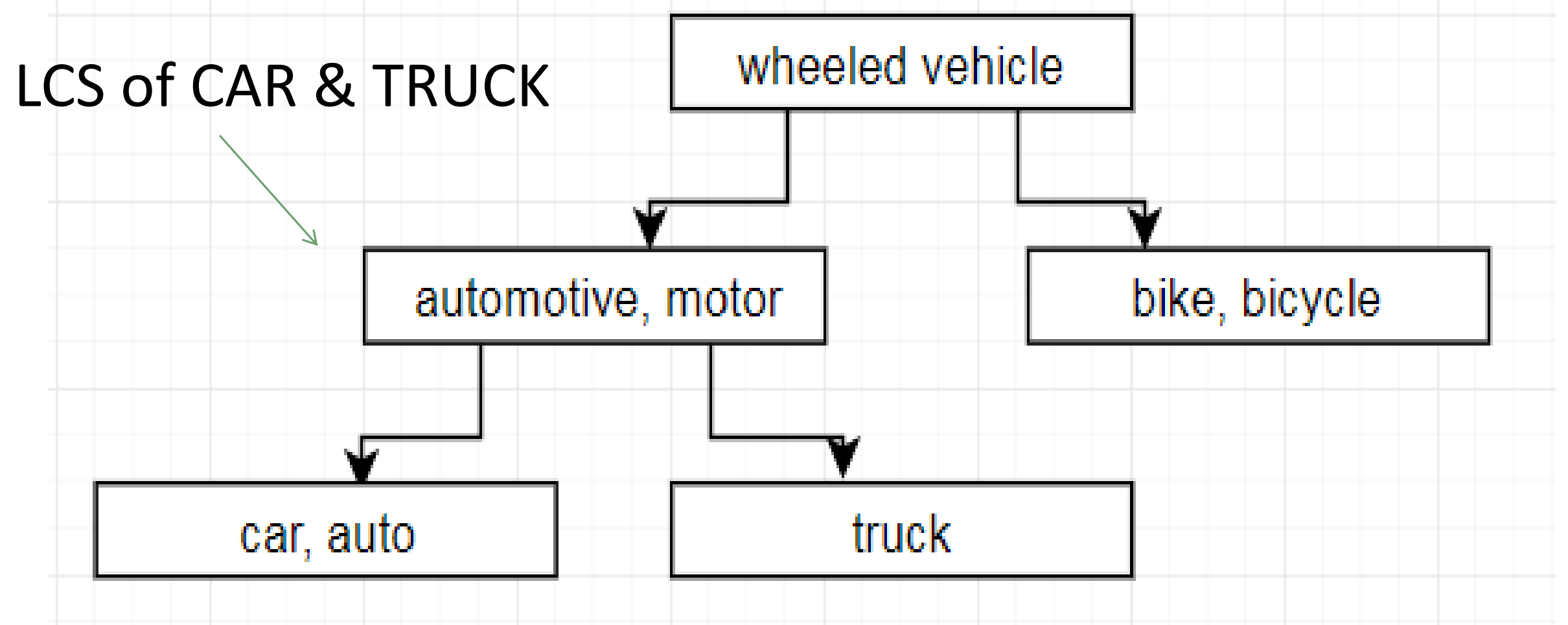
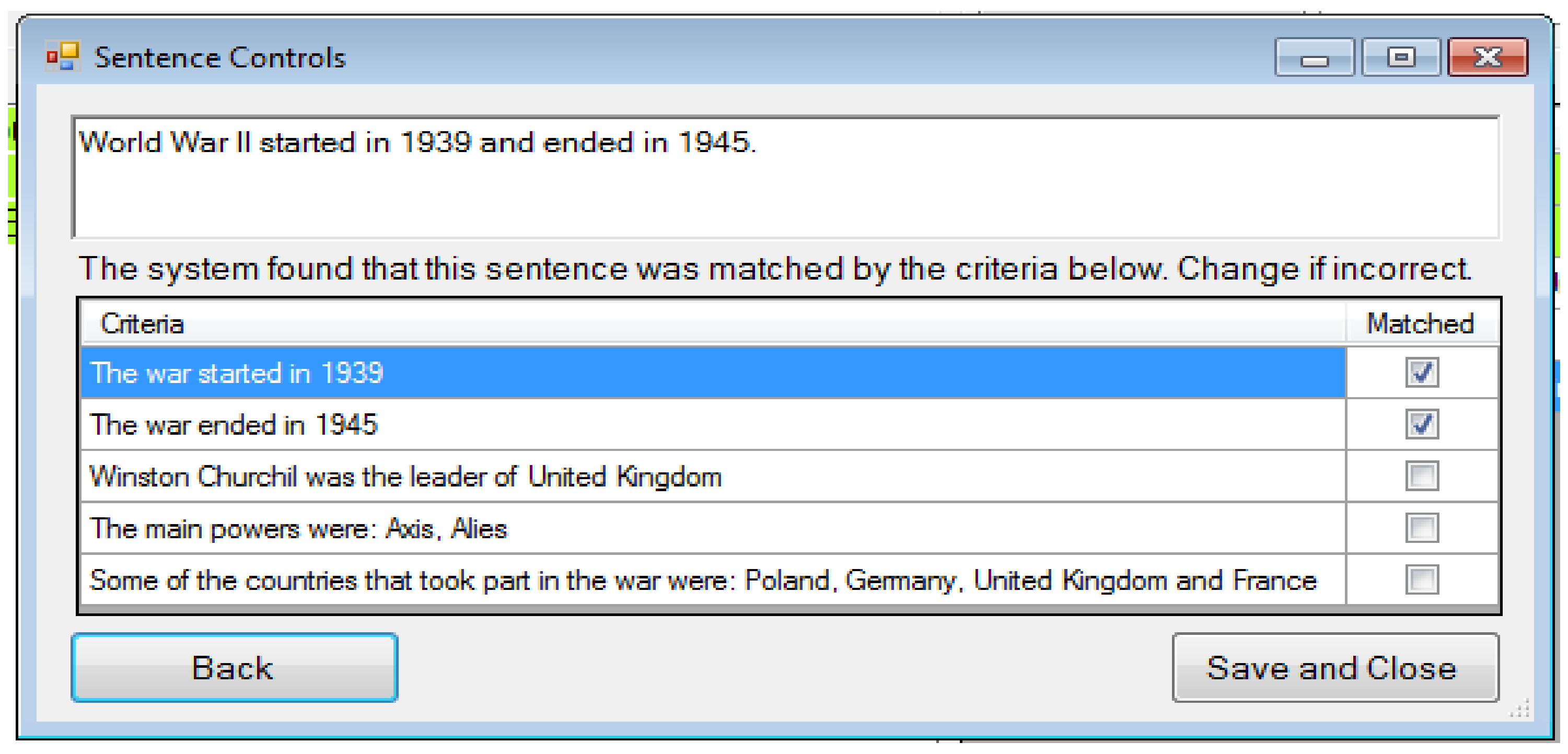


Figure 3. Synset of the word car.

Each word is then paired with its most similar corresponding word and an average of all tokens is calculated. If the value is greater than the set threshold, the matched sentence and criteria are shown to the teacher.

Results

The developed prototype showed great results in; successfully tagging the words with the correct types of speech - 92% accuracy across various data sets (Honnibal, 2013); calculating the semantic similarity between synonyms – 74% accuracy (Silmani, 2013). Word Sense Disambiguation however, has only been rated with less than 40% accuracy (Banerjee and Penderson, 2002).



World War II started in 1939 and ended in 1945.

The system found that this sentence was matched by the criteria below. Change if incorrect.

Criteria	Matched
The war started in 1939	☑
The war ended in 1945	☑
Winston Churchil was the leader of United Kingdom	☐
The main powers were: Axis, Alies	☐
Some of the countries that took part in the war were: Poland, Germany, United Kingdom and France	☐

Figure 4. Screenshot showing matched sentence and criteria.

Conclusion

To conclude, the project dives deep into the existing software solutions that could help speed up the process of marking essays and lower the teachers' work load. Moreover, the developed software explores the field of Natural Language Processing, displaying different techniques and methods that can be used to automate said process. The success of the software however is highly dependant on the success of its algorithms, which as shown above have their shortcomings and require further development.

References

Banerjee, S. and Pedersen, T. (2002) 'An adapted Lesk algorithm for word sense disambiguation using WordNet.' In *International conference on intelligent text processing and computational linguistics,* Duluth USA, 3 February. Springer, Berlin, Heidelberg [Online]. Available at: https://link.springer.com/chapter/10.1007/3-540-45715-1_11 (Accessed: 25 March 2019).

Dikli, S. (2006). 'An overview of automated scoring of essays.' *The Journal of Technology, Learning and Assessment*, *5*(1).

Droogenbroeck, F.V., Spruyt, B. and Vanroelen, C. (2014) '*Burnout among senior teachers: Investigating the role of workload and interpersonal relationships at work', Teaching and Teacher Education,* 43, pp. 99-109.

Gole, S. (2015) *Words similarity/relatedness using WuPalmer Algorithm* Available at: https://blog.thedigitalgroup.com/words-similarityrelatedness-using-wupalmer-algorithm (Accessed: 22 March 2019).

Honnibal, M. (2013) *A Good Part-of-Speech Tagger in about 200 Lines of Python* Available at: https://explosion.ai/blog/part-of-speech-pos-tagger-in-python (Accessed: 04 March 2019).

Lesk, M. (1986) 'Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone.' In *Proceedings of the 5th annual international conference on Systems documentation* Morristown, June. Available at: http://promethee.philo.ulg.ac.be/engdep1/download/prolog/lexdis/docs/lexdis/otherpap/Lesk%20clean.pdf (Accessed: 21 March 2019).

National Education Union (2014) *Teachers and Workload*. Available at: https://neu.org.uk/sites/neu.org.uk/files/files/NEU_Workload_Survey_Report_March_2018.pdf (Accessed: 15 November 2018).

Slimani, T. (2013) 'Description and evaluation of semantic similarity measures approaches.' *International Journal of Computer Applications*, 80(10), pp. 25-33.

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com