**ISA 562100 NLP Lab -- Project Description**

Jason S. Chang

2018/05/08

**Term Project**

The purpose of the research project is for the students to learn how to formulate an NLP problem/task/application and to experience how to solve it using methods, algorithms and techniques taught in class. Students are encouraged to identify new problems/tasks/applications. However, we also provide students with default topics, if they can not come up with a topic.

Building a demo is a must. The students will conduct experimental evaluation on an interesting dataset and will analyze the obtained results. We suggest that the students to work in groups of 2 or more people.

**Project Timeline**

2018/0508: Term Project Announcement and forming teams + discussion and proposal drafting

2016/0515: Lab work: Identify mis-collocations

2016/0522: Interim oral and written report

2016/0529: Lab work: Search engine and user interface

2016/0605: Group report and discussion with TAs

2016/0612: Final report due

**Sample Project**

* **Grammatical Error Correction based on Machine Translation** 
  + 輸入一句錯誤英文，翻譯成為正確英文
  + 利用WikEd Error Corpus產生「翻譯片語表」（phrase table）
  + 參考
* **Grammatical Error Correction based on Synchronous Pattern Grammar**
  + 輸入一句錯誤英文，翻譯成為正確英文
  + 利用WikEd Error Corpus產生「同步樣式文法」（Synchronous Pattern Grammar）

**NLPLAB Project Proposal**

English Grammatical Error Correction   
Based on Machine Translation

John Doe 1

John Doe 2

John Doe 3

（寫名字不要寫學號）

1. Problem Statement

利用機器翻譯的decode模型實作英文改錯系統，將改錯視為一種翻譯，由錯的句子改成對的句子。可以使用混合策略：

* 由字詞、詞性構成的改錯規則
* 由 WikEd Error Corpus 擷取的 Synchronous Pattern Grammar, 例如 discuss about something | discuss something 和 loan something to something | lend something to something
* 最後還要考慮 language model 是否有大幅度提昇

1. Resources
2. EFCAMDAT Corpus
3. WikEd Error Corpus
4. British National Corpus
5. Plan

|  |  |  |
| --- | --- | --- |
| Week | Date | Content |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |

1. Evaluation

利用 Precision, Recall和F-score評估模型的效果，公式如下：

*Precision:*

*Recall:*

*F-score:*

1. Reference
2. Grundkiewicz, Roman, and Marcin Junczys-Dowmunt. "The WikEd Error Corpus: A Corpus of Corrective Wikipedia Edits and Its Application to Grammatical Error Correction." *Advances in Natural Language Processing*. Springer International Publishing, 2014. 478-490. <http://emjotde.github.io/publications/pdf/mjd.poltal2014.draft.pdf>
3. Grundkiewicz, Marcin Junczys-Dowmunt Roman. "The AMU System in the CoNLL-2014 Shared Task: Grammatical Error Correction by Data-Intensive and Feature-Rich Statistical Machine Translation." *CoNLL*-2014 (2014): 25.
4. Felice, Mariano, et al. "Grammatical error correction using hybrid systems and type filtering." *CoNLL*-2014 (2014): 15.
5. WikEd Error Corpus. <http://romang.home.amu.edu.pl/wiked/wiked.html>

**NLPLAB Project Proposal**

Finding Synonyms for a Given Word with Part-of-speech and Sense

John Doe 1

John Doe 2

John Doe 3

（寫名字不要寫學號）

1. Problem Statement

我們製作一個系統，能利用 Google 預先訓練好的 word2vec 的詞彙內嵌向量，來計算某一給予字詞（特定詞性、語意）的同義詞。

Google 預先訓練好的 word2vec 的詞彙內嵌向量，可以提供單字（如 run）的同義詞。然而，這些同義詞是所有相關詞性、語意 (如動詞：「跑」、「逃」、「執行」、名詞：「壘」、「賽跑」)的集合。

|  |
| --- |
|  |
|  |
|  |
| >>> pprint([x for x in model.most\_similar(positive=['run'], topn=50) if x[0] in wndict ])  [('runs', 0.65), ('running', 0.61), ('drive', 0.48), ('ran', 0.48), ('scamper', 0.47), ('go', 0.46),  ('walk', 0.46), ('sacrifice\_fly', 0.42), ('trot', 0.41), ('solo\_homer', 0.41), ('homer', 0.41),   ('Running', 0.40), ('plated', 0.40), ('Run', 0.39), ('stretch', 0.39)] |

理想的結果，應該不是直接使用Google 預先訓練好的 word2vec。因為，如此的結果會有有幾個問題：

1. 如果取最高相似度的同義詞，通常會集中在比較高頻的詞性、語意的同義詞。其他較低頻的詞性、語意的同義詞，就不會出現。

2. 這些同義詞隨機排列，不分詞性、語意，就呈現給使用者。

理想的結果，應該是像辭典、同義詞那樣，將word2vec提供的同義詞，依照詞性、語意，分成一組一組加上標籤，來呈現給使用者。

* **plant, works, industrial plant**: building, facility, construction, factory, project
* **plant, flora, plant life**: vegetation, fauna, wildflower, grasses

由 WikEd Error Corpus 擷取的 Synchronous Pattern Grammar, 例如 discuss about something | discuss something 和 loan something to something | lend something to somethingyi

* 最後還要考慮 language model 是否有大幅度提昇

2. 研究資源

1. EFCAMDAT Corpus
2. WikEd Error Corpus
3. British National Corpus

3. 規劃時程

|  |  |  |
| --- | --- | --- |
| Week | Date | Content |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |

４. 參考文獻

1. Grundkiewicz, Roman, and Marcin Junczys-Dowmunt. "The WikEd Error Corpus: A Corpus of Corrective Wikipedia Edits and Its Application to Grammatical Error Correction." *Advances in Natural Language Processing*. Springer International Publishing, 2014. 478-490. <http://emjotde.github.io/publications/pdf/mjd.poltal2014.draft.pdf>
2. Grundkiewicz, Marcin Junczys-Dowmunt Roman. "The AMU System in the CoNLL-2014 Shared Task: Grammatical Error Correction by Data-Intensive and Feature-Rich Statistical Machine Translation." *CoNLL*-2014 (2014): 25.
3. Felice, Mariano, et al. "Grammatical error correction using hybrid systems and type filtering." *CoNLL*-2014 (2014): 15.
4. WikEd Error Corpus. <http://romang.home.amu.edu.pl/wiked/wiked.html>