Detecting Learner Errors Using Compositional Distributional Semantics

Cambridge ALTA

University of Cambridge

{ Ekaterina Kochmar, Ted Briscoe }   The ALTA Institute, Computer Laboratory

Introduction

The number of non-native speakers of English is growing every year, and **automated learner error detection and correction** has recently become a popular application area for **machine learning** (**ML**) algorithms in natural language processing. Most previous research focuses on function words and casts the task as a multi-class classification problem. In our research, we look at error detection and correction for more challenging errors in **content words** and investigate how ML algorithms can be applied.

Objectives

The focus and objectives of this research:

1. We automatically detect and correct **learner** errors in written English
2. We investigate errors in the choice of **content words**: *adjectives*, *nouns* and *verbs*
3. We take the meaning into account → use **compositional distributional semantics**
4. We use machine learning (**ML**) algorithms to detect and correct errors

Semantic Approaches

	bloom	buy	garden	grow	tall	...
rose	25	18	20	33	8	...
flower	34	23	30	38	10	...
house	0	40	24	5	21	...

Figure 1: Distributional profiles

♦ **Distributional approach**: *"You shall know a word by the company it keeps"* (Firth)
We collect the word co-occurrences from data and build semantic vectors for words within combinations. Distributions capture word meaning.

♦ **Compositional approach**: we create word combination vectors via composition of word vectors.

- $(blue\_rose)_i = blue_i + rose_i$
- $(blue\_rose)_i = blue_i \times rose_i$

♦ **Features**: extract features that describe the differences between the vectors for the correct and incorrect combinations:

- *vector length*
- *distance/cosine to input words*
- *density of the neighbourhoods*
- *overlap between the neighbours for the combinations and for the input words*

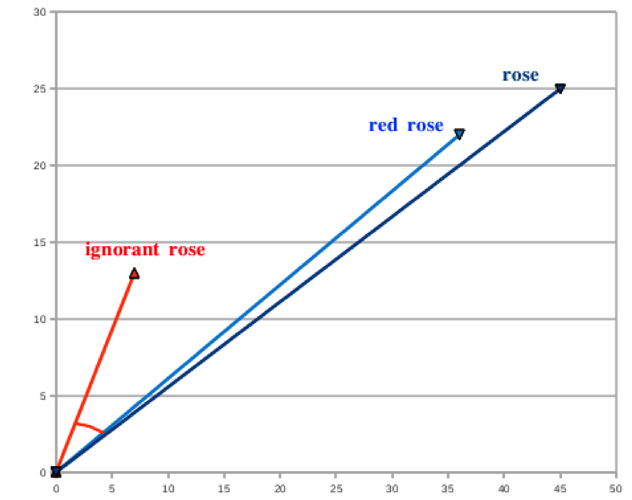


Figure 2: Distance to the input noun

Data & Methods

♦ **Data** is extracted from the *Cambridge Learner Corpus* (CLC), and contains texts written by non-native English speakers with the examples of the correctly as well as incorrectly chosen words.
♦ **The task** is to automatically distinguish between the two classes.
♦ **Previous research** has cast the task as multi-class classification, but focused on predefined set of classes (= number of potential corrections).
♦ **Challenges for content words**:

- How many classes (e.g., as many as there are adjectives in English)?
- Corrections depend on the original word:
  **big history* vs *long history*
  **big conversation* vs *long conversation*
  vs *serious conversation*
- Confusions are caused by different reasons:
  **big anger* vs *great anger* [meaning]
  **classic dance* vs *classical dance* [form]

♦ **Method**: treat as *binary* classification (correct vs. incorrect); encode semantics in the features

ML for Error Detection

Features encode properties of semantic vectors. We use *Decision Tree* classifier with feature value binning.

Combinations	Accuracy	LB	UB
$AN_{-context}$	0.8113	0.7889	0.8650
$AN_{+context}$	0.6535	0.5084	0.7467
$VN_{-context}$	0.6577	0.5557	0.8217
$VN_{+context}$	0.6491	0.6086	0.8467

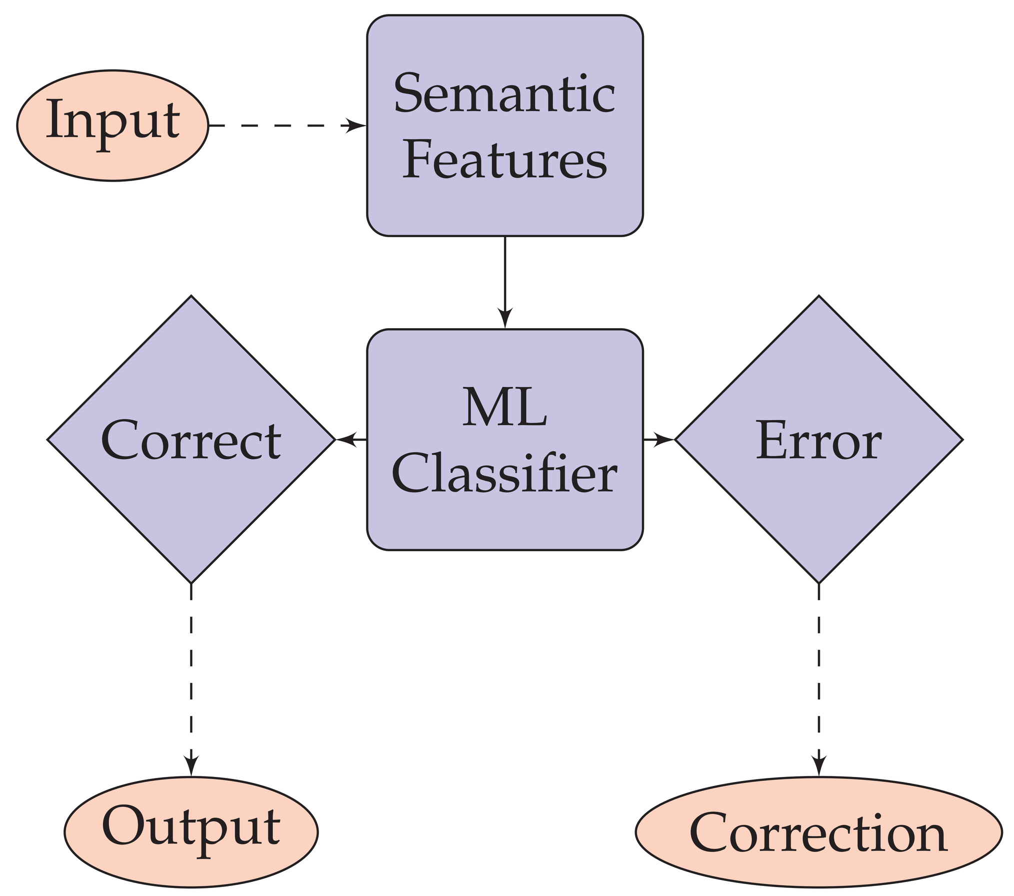
Table 1: Results

LB = *lower bound*, majority class distribution
UB = *upper bound*, inter-annotator agreement

Combinations	Precision	Recall	F_1
$AN_{-context}$	0.8193	0.9762	0.8909
$AN_{+context}$	0.7500	0.2488	0.3736
$VN_{-context}$	0.6173	0.7226	0.6558
$VN_{+context}$	0.7071	0.5898	0.6409

Table 2: Precision, recall and F_1

Conclusion



- We have showed that our algorithm detects errors with high accuracy (close to *UB*)
- There is still some room for improvement
- The features derived using semantics and capturing word meaning are useful
- The algorithm shows high precision → it is reliable in practice
- Major source of misclassification – cases where confusion occurs due to similarity in meaning:
  **small speech* vs *short speech*
  **rise punctuality* vs *increase punctuality*

References

[1] E. Kochmar and T. Briscoe. Detecting Learner Errors in the Choice of Content Words Using Compositional Distributional Semantics. 2014.

[2] E. Kochmar and T. Briscoe. Capturing Anomalies in the Choice of Content Words in Compositional Distributional Semantic Space. 2013.

Future Research

There is an increasing need in error detection and correction algorithms for non-native speakers and writers. We plan to extend current research investigating *error types* other than those currently addressed, *wider use of context* (e.g., via topic modelling), *feature engineering* and other feature types (e.g., neural network language models currently applied), and other *machine learning algorithms*. The next step is to apply an *error correction* algorithm to the errors identified.

Contact Information

Web www.cl.cam.ac.uk/~ek358/
Email Ekaterina.Kochmar@cl.cam.ac.uk

Data ilexir.co.uk/media/an-dataset.xml