SPELL CORRECTION

Cmpe561 Research Project

Mehmet Kose, Utkan Gezer

Outline

1. Introduction
2. Baseline
3. Context Words
4. Collocations
5. Decision Lists
6. Bayesian Classifiers
7. A Combined Approach
8. Other Languages: Arabic
9. Other Languages: German

Introduction

- The problem of spell correction is formulated as: given an alphabet Σ, a dictionary D consisting of strings in Σ^*, and a spelling error s where $s \notin D$ and $s \in \Sigma^*$, find the correction $c \in D$, so that c is most likely to have been erroneously typed as s [1]
- Context insensitive

Context-sensitive Spell Correction

- Confusion sets
- $C = \{w_1, \ldots, w_n\}$ where each word w_i is ambiguous with each other word in the set
- If $C = \{desert, dessert\}$, whenever the spell correction program sees an occurrence of either *desert* or *dessert*, it takes it to be ambiguous and tries to infer from the context which of the two it should be

Ways to obtain confusion sets

- Finding words in the dictionary that are one typo away from each other
- Finding words that have the same or similar pronunciation
- Lists of words that are commonly confused relying on statistics

Baseline Method

- Sets the lowest standard
- During disambiguation of words w_1 through w_n in the confusion set, it ignores the context and always favors the statistically most commonword
- If C = {*desert*, *dessert*} and *desert* occurs more often than *dessert* in the training set, all occurrences of *desert* should be left as it is and of *dessert* should be changed to *desert*

Confusion set	No. of training cases	No. of test cases	Most frequent word	Baseline
whether, weather	331	245	whether	0.922
I, me	6125	840	I	0.886
its, it's	1951	3575	its	0.863
past, passed	385	397	past	0.861
than, then	2949	1659	than	0.807
being, begin	727	449	being	0.780
effect, affect	228	162	effect	0.741
your, you're	1047	212	your	0.726
number, amount	588	429	number	0.627
council, counsel	82	83	council	0.614
rise, raise	139	301	rise	0.575
between, among	1003	730	between	0.538
led, lead	226	219	led	0.530
except, accept	232	95	except	0.442
peace, piece	310	61	peace	0.393
there, their, they're	5026	2187	there	0.306
principle, principal	184	69	principle	0.290
sight, site, cite	149	44	sight	0.114

Context Words

- The identity of an ambiguous word can be extracted from the words around it
- If the target word is unclear to be either *desert* or *dessert*, and there are words like *arid*, *sand*, *sun*, then it's most likely *desert*
- The probability for each wi is calculated using Bayes' rule [2]:

$$p(w_i|c_{-k},\ldots,c_{-1},c_1,\ldots,c_k) = \frac{p(c_{-k},\ldots,c_{-1},c_1,\ldots,c_k|w_i)p(w_i)}{p(c_{-k},\ldots,c_{-1},c_1,\ldots,c_k)}$$

- Due to sparse data problem, we assume that the presence of a word is independent of the presence of any other word. This turns the previous equation into the following:

$$p(c_{-k}, \ldots, c_{-1}, c_1, \ldots, c_k | w_i) = \prod_{j \in -k,\ldots,-1,1,\ldots,k} p(c_j | w_i)$$

Confusion set	Baseline	Cwords ±3	Cwords ±6	Cwords ±12	Cwords ±24
whether	0.922	0.902	0.922	0.927	0.922
I	0.886	0.914	0.893	0.883	0.851
its	0.863	0.862	0.795	0.743	0.702
past	0.861	0.861	0.849	0.801	0.743
than	0.807	0.931	0.901	0.896	0.855
being	0.780	0.791	0.795	0.793	0.755
effect	0.741	0.747	0.741	0.759	0.716
your	0.726	0.816	0.783	0.774	0.736
number	0.627	0.646	0.622	0.636	0.639
council	0.614	0.639	0.614	0.602	0.614
rise	0.575	0.575	0.575	0.585	0.498
between	0.538	0.759	0.697	0.671	0.586
led	0.530	0.530	0.530	0.521	0.557
except	0.442	0.695	0.526	0.516	0.558
peace	0.393	0.754	0.705	0.574	0.574
there	0.306	0.726	0.623	0.557	0.466
principle	0.290	0.290	0.290	0.290	0.435
sight	0.114	0.455	0.250	0.364	0.318
Avg no. of context words		27.9	36.9	55.9	92.9

Collocations

- The method of context words is good at capturing generalities that depend on the presence of nearby words
- But it ignores their order
- A collocation expresses a pattern of syntactic elements around the target word, such as words and part-of-speech tags

- If we consider the example of *desert*, *dessert*, a collocation for *desert* might be
  - `PREP` *the* ____
- This collocation would match the sentences:
  - Travelers entering from the *desert* were confounded . . .
  - . . . along with some guerrilla fighting in the *desert*
  - . . . two ladies who lay beside him in the *desert* . . .

- Unlike context words, collocations cannot be assumed to be independent.
- Consider the following collocations for *desert*:
  - `PREP` *the* ___
  - *in the* ___
  - *the* ___
- These collocations are highly interdependent (they *conflict*)
- If two pieces of evidence conflict, one of them is eliminated the decision is based on the rest of the evidences
- The most common approach to elimination is to assign each evidence a *strength* and eliminate the one with lower strength

Decision Lists

- The method of decision lists is a hybrid method that combines context words and collocations [3]
- Context words pick up generalities that are expressed in an order-independent way whereas collocations capture order-dependent generalities
- One big list of all features
- The features are sorted in order of decreasing strength
- The first feature that matches is used to classify the target word

Bayesian Classifiers

- Decision lists prove that combining two complementary methods -context words and collocations- result in effective results
- Uses single strongest piece of evidence for a given problem
- Golding et al. [4] propose that a better performance can be obtained by taking into account all available evidence
- It traverses the entire list, combining evidence from all matching feature
- Resolve conflicts if they arise

Confusion set	Baseline	Cwords ± 3	Collocs ≤ 2	Dlist Rely	Bayes Rely	Trigrams
whether	0.922	0.902	0.931	0.935	0.935	0.873
I	0.886	0.914	0.981	0.980	0.985	0.985
its	0.863	0.862	0.945	0.931	0.942	0.965
past	0.861	0.861	0.909	0.932	0.924	0.955
than	0.807	0.931	0.965	0.967	0.973	0.780
being	0.780	0.791	0.853	0.842	0.869	0.978
effect	0.741	0.747	0.821	0.821	0.827	0.975
your	0.726	0.816	0.887	0.868	0.901	0.958
number	0.627	0.646	0.646	0.629	0.662	0.636
council	0.614	0.639	0.639	0.627	0.639	0.651
rise	0.575	0.575	0.807	0.804	0.807	0.574
between	0.538	0.759	0.730	0.659	0.786	0.538
led	0.530	0.530	0.840	0.840	0.840	0.909
except	0.442	0.695	0.789	0.789	0.811	0.695
peace	0.393	0.754	0.869	0.852	0.852	0.393
there	0.306	0.726	0.932	0.914	0.916	0.961
principle	0.290	0.290	0.812	0.812	0.812	0.609
sight	0.114	0.455	0.318	0.432	0.455	0.250

A Combined Approach: Trigrams and Bayesians [6]

- For a given confusable word in a sentence, the most likely part-of-speech is determined for that location using trigrams of POS.
- If there is only one word that can be tagged with that POS among the confusion set of the confusable word, then we conclude.
- Multiple words? Then the features are taken into account using the bayesian classifiers to pick one from those words.

Confusion set	System scores			
	Base	T	B	TB
their, there, they're	56.8	97.6	94.4	97.6
than, then	63.4	94.9	93.2	94.9
its, it's	91.3	98.1	95.9	98.1
your, you're	89.3	98.9	89.8	98.9
begin, being	93.2	97.3	91.8	97.3
passed, past	68.9	95.9	89.2	95.9
quiet, quite	83.3	95.5	89.4	95.5
weather, whether	86.9	93.4	96.7	93.4
accept, except	70.0	82.0	88.0	82.0
lead, led	46.9	83.7	79.6	83.7
cite, sight, site	64.7	70.6	73.5	70.6
principal, principle	58.8	88.2	85.3	88.2
raise, rise	64.1	64.1	74.4	76.9
affect, effect	91.8	93.9	95.9	95.9
peace, piece	44.0	44.0	90.0	90.0
country, county	91.9	91.9	85.5	85.5
amount, number	71.5	73.2	82.9	82.9
among, between	71.5	71.5	75.3	75.3

Table 1: Overall performance of all methods: Baseline (Base), part-of-speech Trigrams (T), Bayes (B), and the combination, Tribayes (TB). System scores are given as percentages of correct predictions[5].

A Combined Approach: Trigrams and Bayesians

- Each one of the individual methods perform worse than their combination.
- Main idea: Complementarity of the methods is utilized.

Confusion set	Different tags				Same tags			
	Break-down	System scores			Break-down	System scores		
		Base	T	B		Base	T	B
their, there, they're	100	56.8	97.6	94.4	0	—	—	—
than, then	100	63.4	94.9	93.2	0	—	—	—
its, it's	100	91.3	98.1	95.9	0	—	—	—
your, you're	100	89.3	98.9	89.8	0	—	—	—
begin, being	100	93.2	97.3	91.8	0	—	—	—
passed, past	100	68.9	95.9	89.2	0	—	—	—
quiet, quite	100	83.3	95.5	89.4	0	—	—	—
weather, whether	100	86.9	93.4	96.7	0	—	—	—
accept, except	100	70.0	82.0	88.0	0	—	—	—
lead, led	100	46.9	83.7	79.6	0	—	—	—
cite, sight, site	100	64.7	70.6	73.5	0	—	—	—
principal, principle	29	0.0	100.0	70.0	71	83.3	83.3	91.7
raise, rise	8	100.0	100.0	100.0	92	61.1	61.1	72.2
affect, effect	6	100.0	100.0	66.7	94	91.3	93.5	97.8
peace, piece	2	0.0	100.0	100.0	98	44.9	42.9	89.8
country, county	0	—	—	—	100	91.9	91.9	85.5
amount, number	0	—	—	—	100	71.5	73.2	82.9
among, between	0	—	—	—	100	71.5	71.5	75.3

Table 2: Performance of the component methods, Baseline (Base), Trigrams (T), and Bayes (B). System scores are given as percentages of correct predictions. The results are broken down by whether or not all words in the confusion set would have the same tagging when substituted into the target sentence. The "Breakdown" columns show the percentage of examples that fall under each condition.

Arabic

- Arabic has a rich and complex morphology as it applies both concatenative and non-concatenative morphotactis
- شَكَرَ (to thank)
- وَشَكَرَتهُ (and she thanked him)
- وَسَيُستَدُعَونَ (and they will be summoned)
- A verb, such as شَكَرَ, generates 2552 valid forms
- A noun, such as مُعَلِّم, generates 519 valid forms

- Attia et al [5] developed a hybrid spell checker for Arabic that has three components which are the following:
  - Error detection through a dictionary (or a reference word list)
  - Candidate generation through edit distance as implemented in a finite state compiler
  - Best candidate selection using an n-gram language model

	Accuracy	Recall	Precision	f-measure
Ayaspell for Hunspell v. 3.4	95.74	96.69	98.26	97.47
Microsoft Word 13	97.68	**99.14**	98.14	98.64
Google Docs (April 2014)	87.91	96.02	90.33	93.09
AraComLex Extended 1.5	**98.63**	99.09	**99.30**	**99.19**

German

- Famously known for allowing the concatenated combinations of words, also known as “compounding”.
- The amount of the valid words is therefore practically indefinite.
- A compound word in German language is not necessarily a plain concatenation of two words, but may rather involve [7]:
  - the addition of a “linking element” in between the segments,
  - the shortening of segments,
  - or a segment getting morphed.
- Complexity of the problem of spell checking in compounding languages is higher.

References

1. E. Brill and R. C. Moore, "An improved error model for noisy channel spelling correction," in Proceedings of the 38th Annual Meeting on Association for Computational Linguistics , pp. 286-293, Association for Computational Linguistics, 2000.
2. W. A. Gale, K. W. Church, and D. Yarowsky, "Discrimination decisions for 100,000-dimensional spaces," in Current Issues in Computational Linguistics: In Honour of Don Walker , pp. 429-450, Springer, 1994.
3. D. Yarowsky, "A comparison of corpus-based techniques for restoring accents in spanish and french text," in Natural language processing using very large corpora , pp. 99-120, Springer, 1999.

4. A. R. Golding, "A bayesian hybrid method for context-sensitive spelling correction," arXiv preprint cmp-lg/9606001 , 1996.
5. M. Attia, P. Pecina, Y. Samih, K. Shaalan, and J. Van Genabith, "Arabic spelling error detection and correction," Natural Language Engineering, vol. 22, no. 5, pp. 751-773, 2016.
6. A. R. Golding and Y. Schabes, “Combining trigram-based and feature-based methods for context-sensitive spelling correction,” in Proceedings of the 34th annual meeting on Association for Computational Linguistics, pp. 71–78, Association for Computational Linguistics, 1996.
7. A. M. Jessee, M. R. Eckert, and K. R. Powell, “Compound word breaker and spell checker,” Nov. 4 2008. US Patent 7,447,627.