

Appendix - Charmanteau: Character Embedding Models For Portmanteau Creation

Varun Gangal *, Harsh Jhamtani *, Graham Neubig, Eduard Hovy, Eric Nyberg

Language Technologies Institute,

Carnegie Mellon University

{vgangal, jharsh, gneubig, hovy, ehv}@cs.cmu.edu

Abstract

This document contains additional information to complement the descriptions and findings reported in the paper, as well as a guide to the supplementary material.

1 Qualitative Examples

In Table 1, we present model outputs for a list of some randomly selected examples from D_{Blind} .

2 Dataset Statistics

Average character length for first word across D_{Large} is 6.41 with variance of 6.51, while for second word is 7.32 with variance of 4.98. Average character length for first word across D_{Wiki} is 6.24 with variance of 5.89, while for second word is 7.24 with variance of 4.56.

3 Output generation - SCORE decoding explanation

The SCORE decoding strategy is described in greater detail here than in the paper, due to paucity of space in the former. This strategy consists of two steps.

1. Candidate Generation

This step takes every non-empty prefix of the first word $x^{(1)}$ and every non-empty suffix of the second word $x^{(2)}$ and generates a candidate list, where each candidate is a distinct prefix-suffix combination. More formally, for the words $x^{(1)}$ and $x^{(2)}$, the list of candidates would be given by all strings of the form $y = \text{concat}(x_{pre}^{(1)}, x_{suf}^{(2)}), x_{pre}^{(1)} \in X_{pre}^{(1)}, x_{suf}^{(2)} \in X_{suf}^{(2)}$, where S_{pre} and S_{suf} are the sets of prefixes and suffixes respectively of a string s . In the case, where $x^{(1)} = \text{car}$

and $x^{(2)} = \text{hijack}$, $X_{pre}^{(1)} = \{c, ca, car\}$ and $X_{suf}^{(2)} = \{k, ck, ack, jack, ijack, hijack\}$

2. Scoring

The candidates are scored using the respective model for $P(y|x)$.

Note that for the GREEDY and BEAM strategies, these two steps are tied together, since we maintain only the k-best (or 1-best, in case of GREEDY) hypotheses after each step.

4 Human Annotation Experiment Details

As mentioned, the study was performed on AMT¹.

3. Word 1: methyl Word 2: alcohol

Now consider following two blended words: methyhol, malcohol

- ☐ methyhol is much better than malcohol
- ☐ methyhol is better than malcohol
- ☐ malcohol is better than methyhol
- ☐ malcohol is much better than methyhol

Figure 1: Example question in a HIT (Human Intelligence Task). Every HIT had 6 such questions.

4.1 Filtering out Unreliable Responses

Since the data collected from the crowdsourced platforms such as Amazon Mechanical Turk can be noisy, we added a *test* question with every HIT (Human Intelligence Task). If the answer to the *test* question is incorrect, we ignore all the responses in that HIT. We use a simple strategy to generate test question: first output is first four characters of the first word of a randomly chosen portmanteau, while the other output is the ground truth. We expect the users to mark ground truth to be *better* or *much better*.

* denotes equal contribution

¹requester.mturk.com

Input	FORWARD	BACKWARD	BASELINE	G.TRUTH
shopping;marathon	shopparathon	shoathon	shon	shopathon
fashion;fascism	fashism	fashism	fashism	fashism
wiki;etiquette	wikiquette	wiquette	wiquette	wikiquette
clown;president	clowident	clownsident	clownt	clownsident
car;hijack	carjack	carjack	cack	carjack
dialectical;materialism	dialerialism	dialerialism	dialism	dialerialism
tinder;disaster	tinter	tindersaster	tindisaster	tindisaster
data;broadcasting	datasting	doadcasting	dating	datacasting
back;acronym	backronym	bacronym	bacronym	backronym
britain;regret	bregret	brigret	bregret	bregret
social;entertainer	socialtainer	sociartainer	sentertainer	socialtainer
chopstick;fork	chopstork	chopfork	chork	chork
happy;harmonius	happonius	happonius	hharmonius	happymonius
flexible;vegetarian	flexarian	flexetarian	flegetarian	flexitarian
laughter;orgasm	lauggasm	laughtergasm	lasm	laughgasm
frequency;recency	frecency	frecency	frecency	frecency
tender;entrepreneur	tenpreneur	tendereneur	tenterpreneur	tenderpreneur
fall;halloween	falloween	falloween	falloween	falloween
frisbee;golf	frolf	frisbolf	frolf	frolf
hitler;hillary	hitlary	hitlery	hitlery	hitlery
trump;economics	trumpics	trumponomics	trumics	trumponomics
flirtation;relationship	flirtionship	flirtationship	flirtationship	flirtationship
next;yesterday	nexterday	nesterday	nexterday	nexterday
lobster;monstrosity	lobstrosity	lonstrosity	lobstrosity	lobstrosity
global;english	glonglish	globlish	glish	globlish
puke;extravaganza	pukaganza	pukaganza	puextravaganza	pukestravaganza
beverage;avalanche	bevalanche	beveranche	bavalanche	bevelanche
excited;dimmer	excimmer	excimmer	excitedimmer	excimmer
phone;amnesia	phonesia	phonesia	phonesia	phonesia
camera;phone	came	camphone	camphone	camphone
bored;ordinary	bordinary	bordinary	bordinary	bordinary
precise;exactly	prexactly	prexactly	prexactly	prezactly

Table 1: Example outputs from different models. Outputs are from best performing configurations of the models. G.TRUTH denotes the ground truth portmanteau.

5 Dataset Specifications

The dataset directory itself contains a *README* which explains its contents. For the sake of clarity, we also provide an explanation here.

5.1 Main Dataset

The dataset file is named *dataset.csv*. The dataset is a comma-separated file, where each line is of the form $(y, x^{(1)}, x^{(2)}, flag)$. y and $x^{(1)}, x^{(2)}$ are the portmanteau and root words respectively. *flag* can have two values, *knight* and *other*. If *flag* = *knight*, the respective example is $\in D_{Wiki}$. If *flag* = *other*, the respective example is $\in D_{Blind}$.

5.2 Baseline Files

We mentioned in our paper that we got the predictions of the trained BASELINE model on the unseen, held-out dataset D_{Blind} (1223 examples) from <http://leps.isi.edu/fst/step-all.php>. For clarity and verifiability, we share these BASELINE results as well. The baseline results are shared as the file *baselineResults.txt*. Each line in the file is space-separated

and is of the form “ $x^{(1)} x^{(2)} y^{BASELINE}$ ”, where $x^{(1)}$ and $x^{(2)}$ are the root words and $y^{BASELINE}$ is the portmanteau predicted by BASELINE.

5.3 Ranklist Files

Additionally, we also submit candidate rank lists from our two best performing approaches - FORWARD+ATTN+INIT+ENSEMBLE+SCORE and BACKWARD+ATTN+INIT+ENSEMBLE+SCORE in Experiment 2, on D_{Blind} . The files are named *best_blind_forward.txt* and *best_blind_backward.txt* respectively. Each file contains a list of examples. For each example, we have a rank-list (one candidate per line) in ascending order of log-loss (lesser loss is better). The top-most element in this list is the word output by our model. Finally the original word (GROUND-TRUTH) and the word output by our model are printed out.

6 Portmanteaus in Some Other Languages - Brief Discussion

Bahasa Indonesia, being a newly constructed official language, has given rise to a wealth of port-

manteaus, especially for governmental and official terminology and acronyms, as explored in greater detail in (Dardjowidjojo, 1979) .

Some languages such as Japanese also have portmanteau-like structures, albeit with fairly regular rules of formation, taking away the need to do any learning. Verifying whether our model can learn these rules automatically, however, could be a potential way of evaluating their goodness.

As a concrete direction for future work, we plan to curate datasets of Japanese and Indonesian portmanteaus, and evaluate our models on them.

References

Soenjono Dardjowidjojo. 1979. Acronymic Patterns in Indonesian. *Pacific Linguistics Series C*, 45:143–160.