ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Tarek Sakakini

Suma Bhat

Pramod Viswanath

IBM

C3SR center for cognitive computing systems research

MORSE:

Semantic-ally Drive-n MORpheme SEgment-er

Samuel MORSE minimized the number of on-off clicks for non-verbal communication.

This MORSE minimizes the vocabulary size for Natural Language Processing systems.

1

Morpheme Segmentation

Morpheme Segmentation

Hopefully

Not a trivial task

+ing

+er

+s

Players

Playing

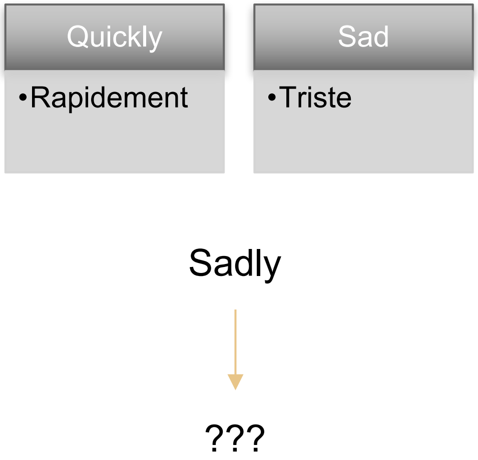
Beijing

Butterflies

Applications

Machine Translation

Model:


Quickly

- Rapidement

Sad

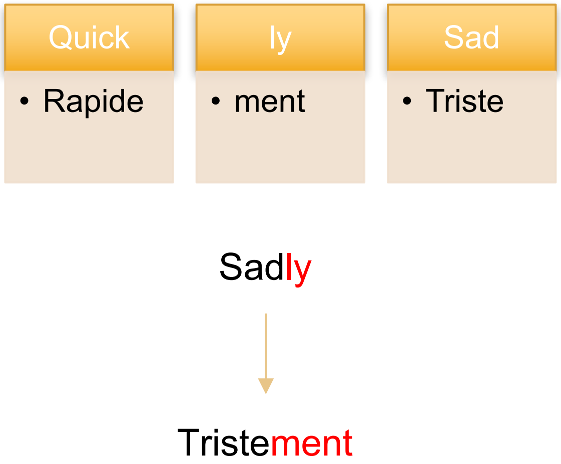
- Triste

Sadly

???

Test:

Model:


Quick

- Rapide

ly

- ment

Sad

- Triste

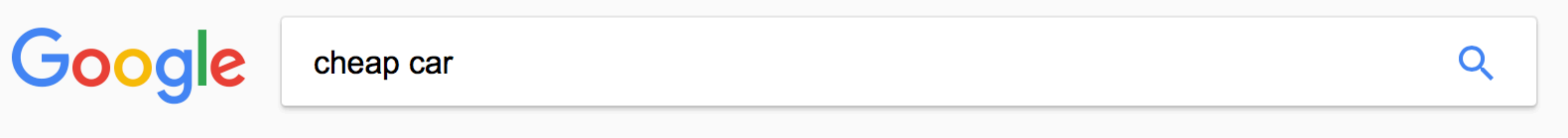
Sadly

Tristement

Test:

Applications

Information Retrieval



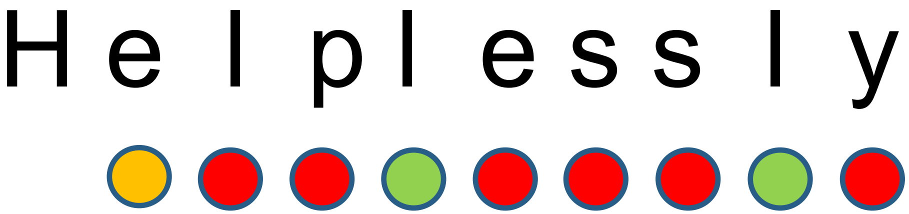
cheap car

Here at Toyota World, we have the **cheap**est **car**s in town.
We are proudly called the first and last stop.
…
…

2

Previous Work

Letter Successor Variety (Harris, 1970)



Helplessly

Morfessor (Creutz and Lagos, 2005)

Help: 2387

Helping: 1586

Helper: 498

Helps: 2437

Jump: 1847

Jumping: 1664

Jumper: 1290

Jumps: 2987

Downsides

Freshman

Butterfl ies ✗

Butterfly ies ✓

Locally Semantic

Cosine similarity

car 🔴 caring

car 🟢 cars

(Schone and Jurafsky, 2000)
(Narasimhan et al., 2015)
(Luo et al., 2017)

Distinguishing criteria

car → cars

player → players

runner → runners

goal → goals

play → plays

car → cars

fine → fines

wheel → wheels

hand → hands

laptop → laptops

lab → labs

3

MORSE

Input:
Word Embeddings

Unsupervised
Morphology Learning

Segmentation:
Optimization Problem

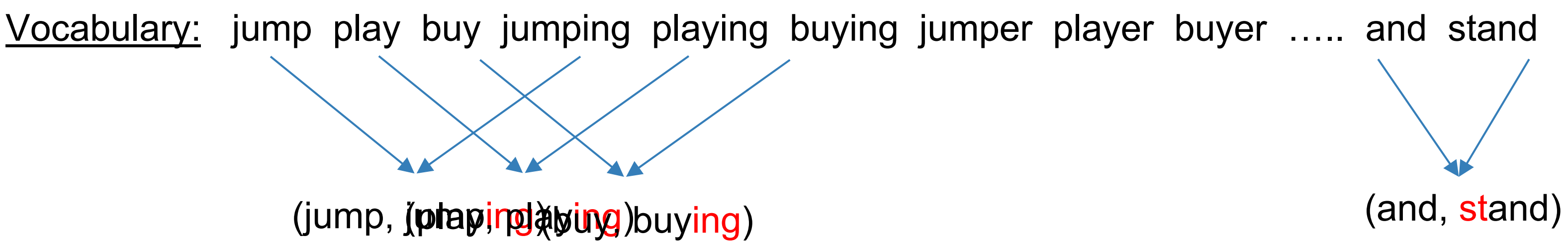
4 hyperparameters:
Small tuning dataset

Step 1

Learning Morphology

(Soricut and Och, 2015)

Collecting candidate morphological rules



Vocabulary: jump play buy jumping playing buying jumper player buyer ….. and stand

(jump, jumping) (play, playing) (buy, buying)

(and, stand)

(suf, ∅, ing):	(jump, jumping)	(play, playing)	(buy, buying)
(suf, ∅, er):	(jump, jumper)	(play, player)	(buy, buyer)
(pre, ∅, st):	(and, stand)	(ore, store)	(one, stone)

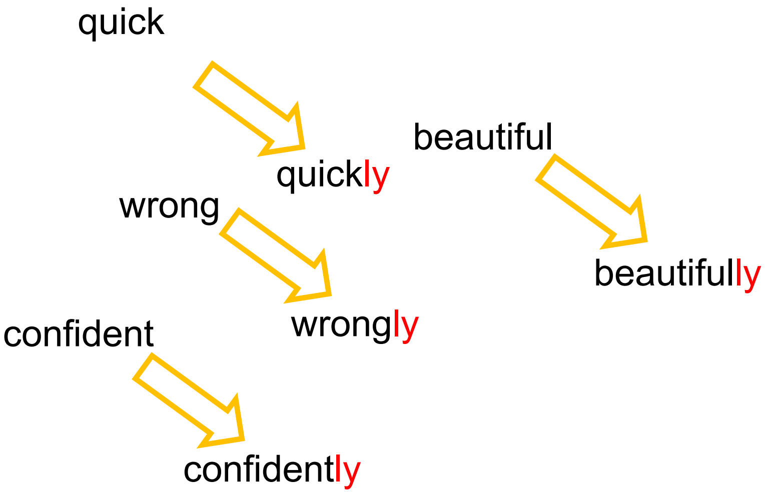
Signals

Orthographic

quick	quickly
beautiful	beautifully
confident	confidently
wrong	wrongly

Semantic

Word Embeddings



quick

quickly

beautiful

beautifully

wrong

wrongly

confident

confidently

What makes a good rule?

Signal 1: Orthography

✓

Rule = (suf, ∅, ly) **Size = 8723**

(quick, quickly) (beautiful, beautifully)

(confident, confidently)

……………………………………………

……………………… (wrong, wrongly)

✗

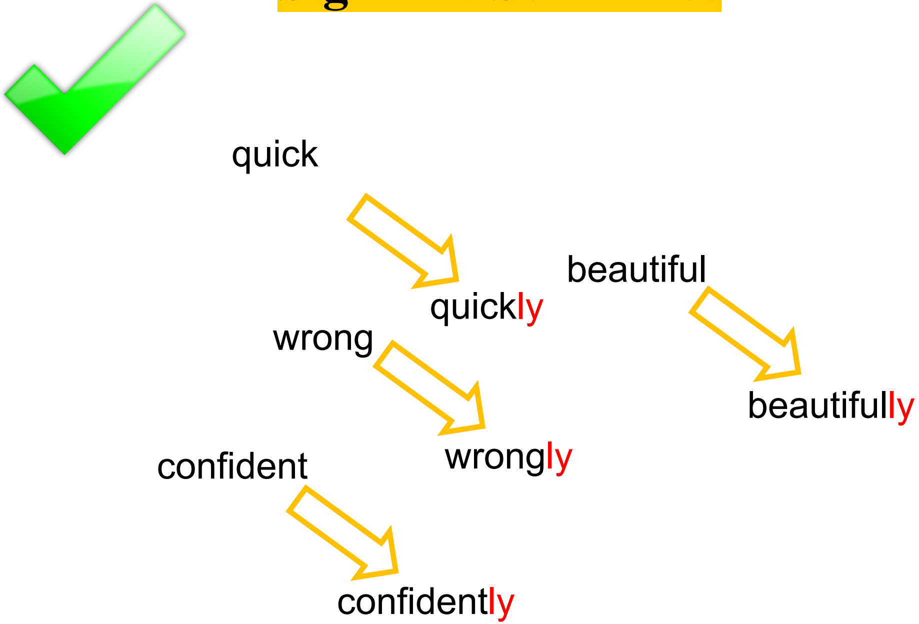
Rule = (pre, ∅, st) **Size= 16**

(ore, store) ……

(amp, stamp)

What makes a good rule?

Signal 2: Semantics



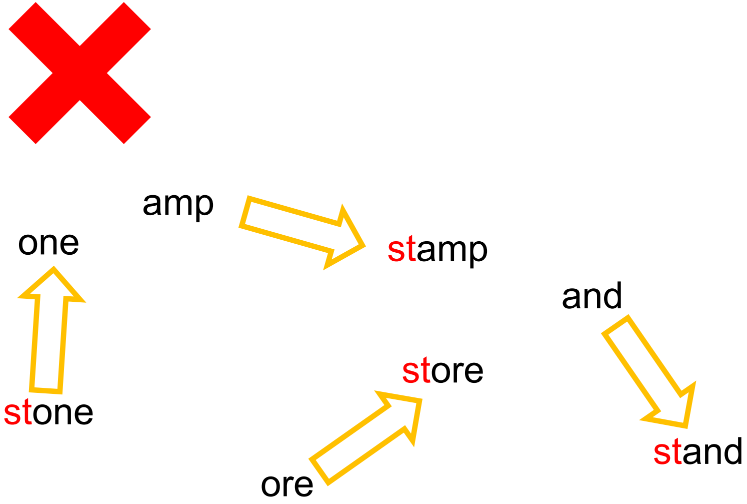
✓

quick → quickly

beautiful → beautifully

wrong → wrongly

confident → confidently



✗

amp → stamp

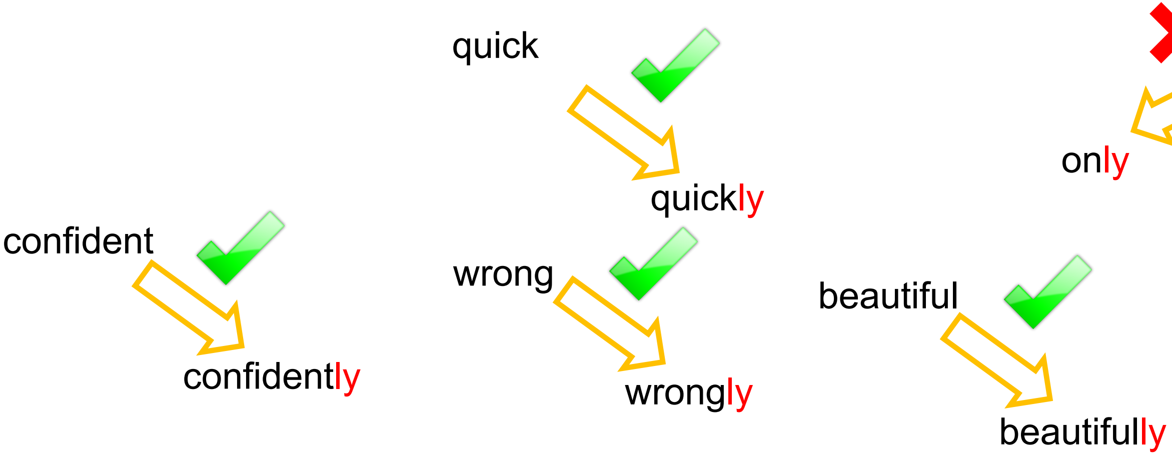
stone → one

and → stand

ore → store

What makes a good member of a rule?

Scope: Vocabulary-Wide



quick ✓

quickly

✗ on

only

confident ✓

confidently

wrong ✓

wrongly

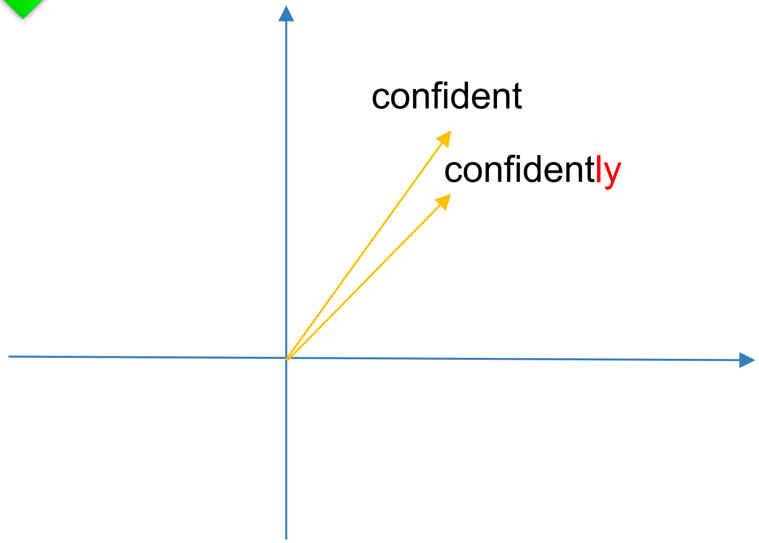
beautiful ✓

beautifully

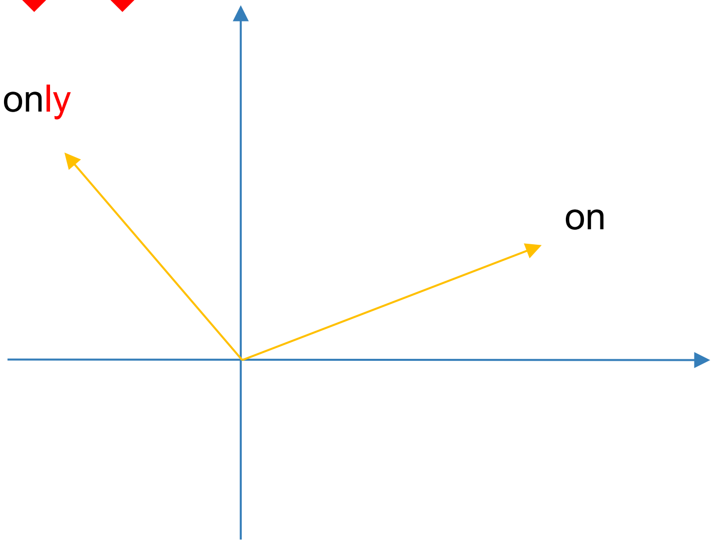
What makes a good member of a rule?

Scope: Local

✓



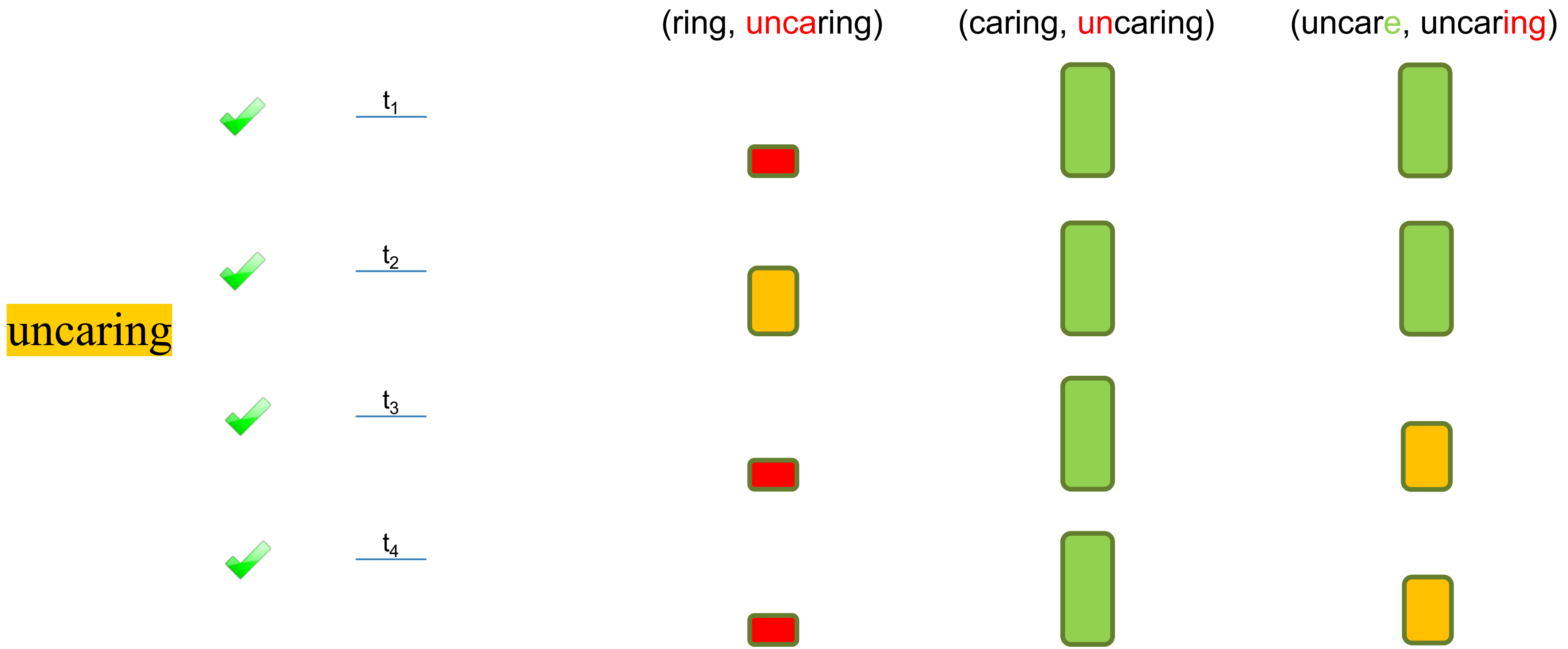
✗



Step 2

Segmenting

Linear Optimization Problem



(ring, uncaring)

(caring, uncaring)

(uncare, uncaring)

uncaring

✓ t_1

✓ t_2

✓ t_3

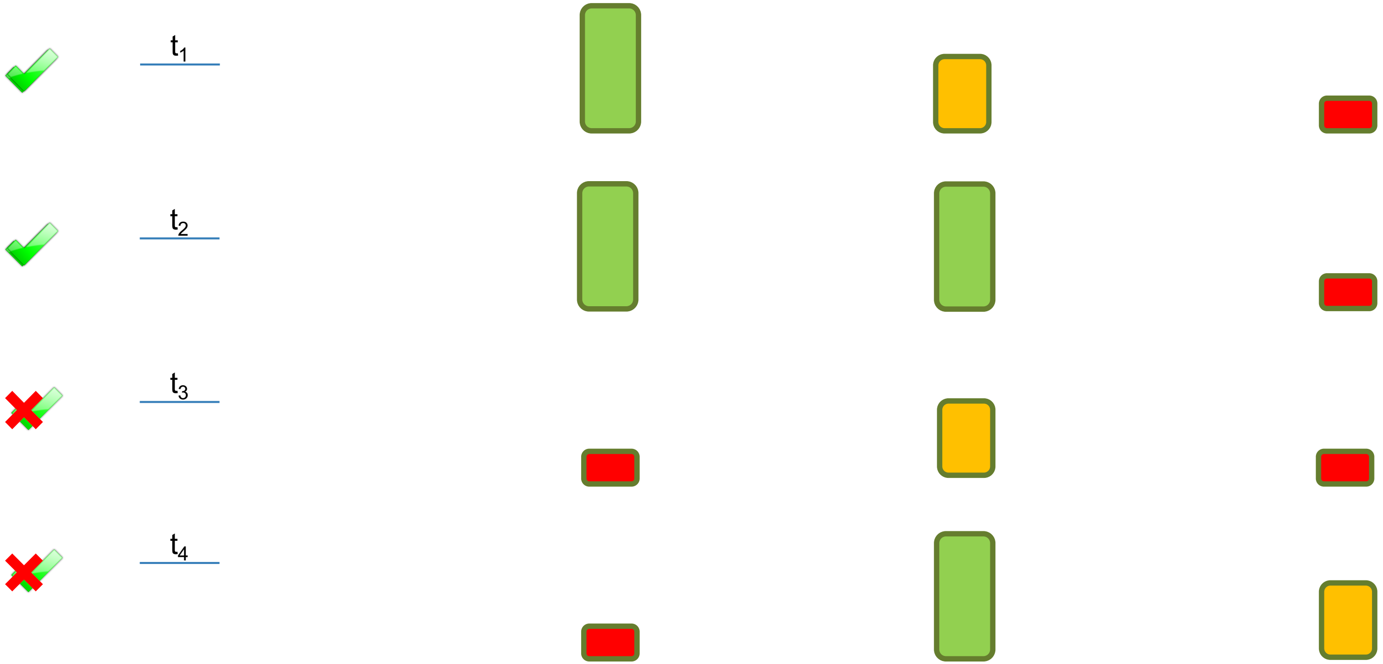
✓ t_4

un + caring

Iterate

caring

(car, caring) (care, caring) (carol, caring)



✓ t_1

✓ t_2

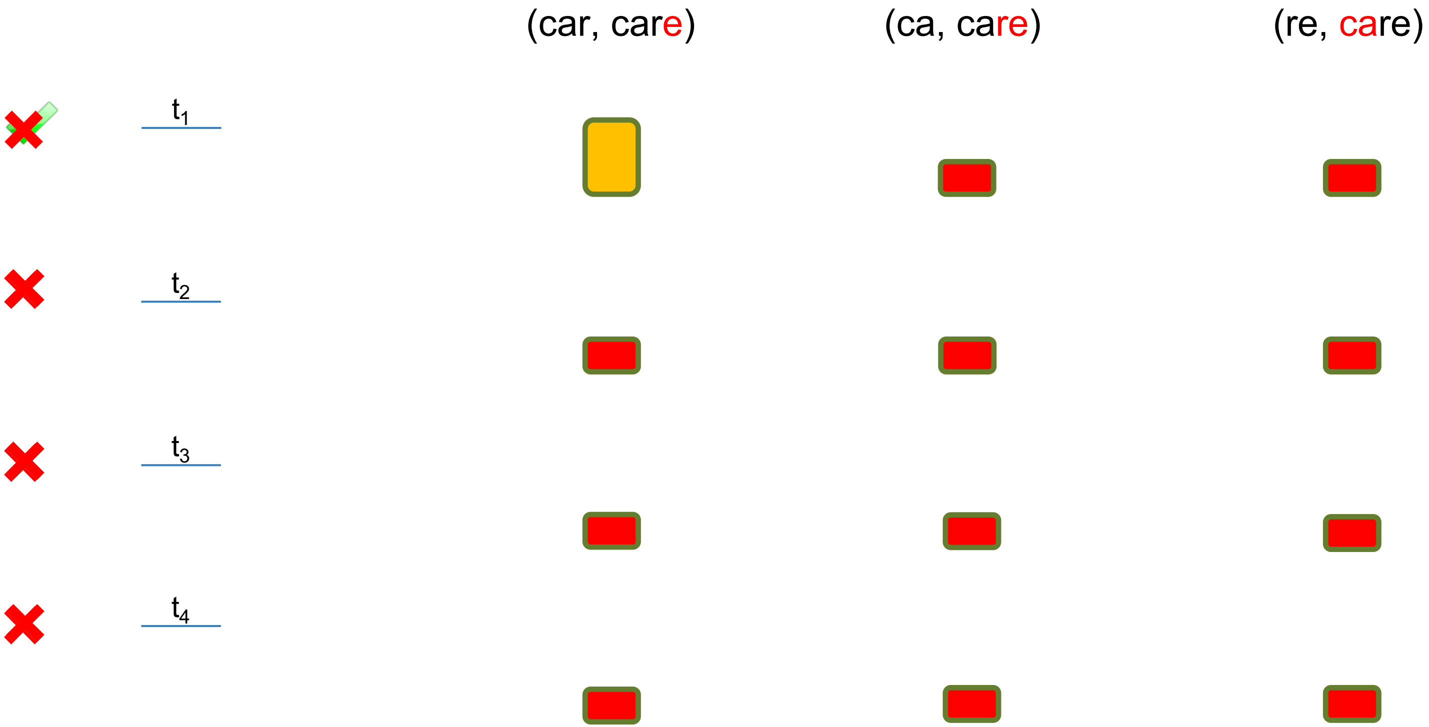
✗ t_3

✗ t_4

un + care + ing

Iterate

care



(car, care)

(ca, care)

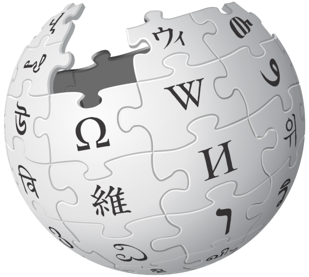
(re, care)

4

Experiments

Experimental Setup

Training



Languages

🇺🇸

🇹🇷

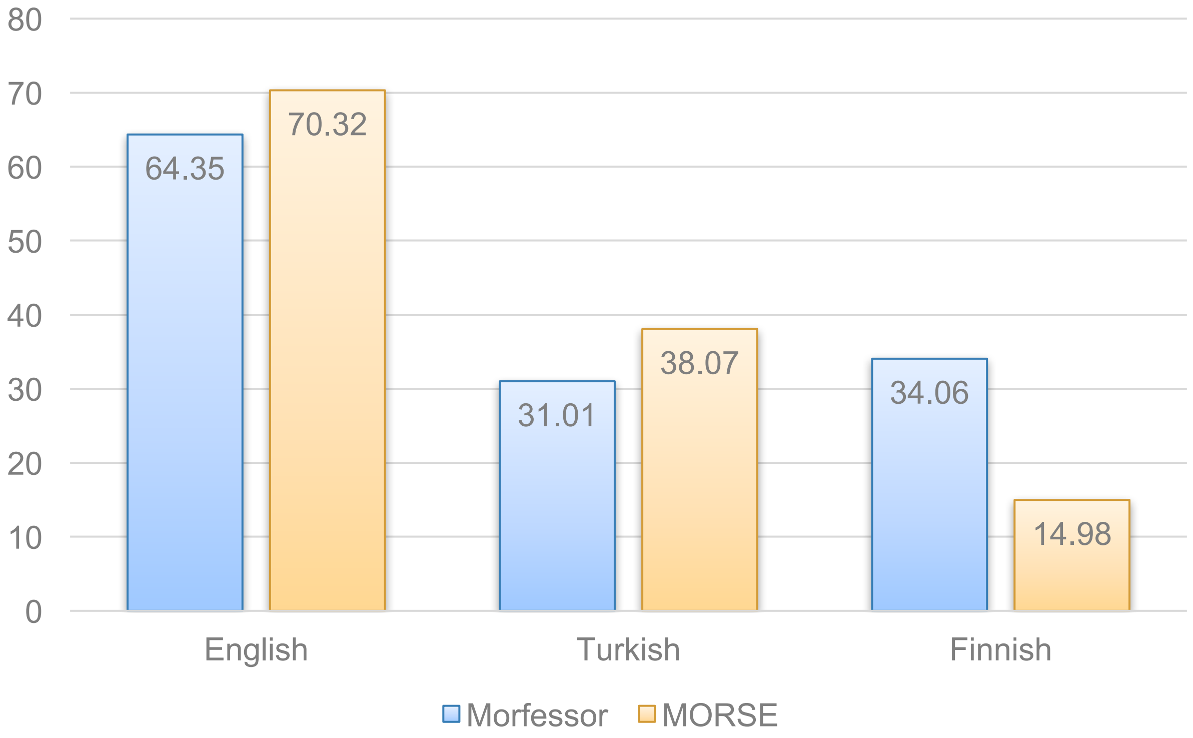
🇫🇮

Gold Datasets

Morpho Challenge

jumping	jump	ing
playing	play	ing
jumps	jump	s
calls	call	s
rooms	room	s

Experiments



Morpho Challenge downsides

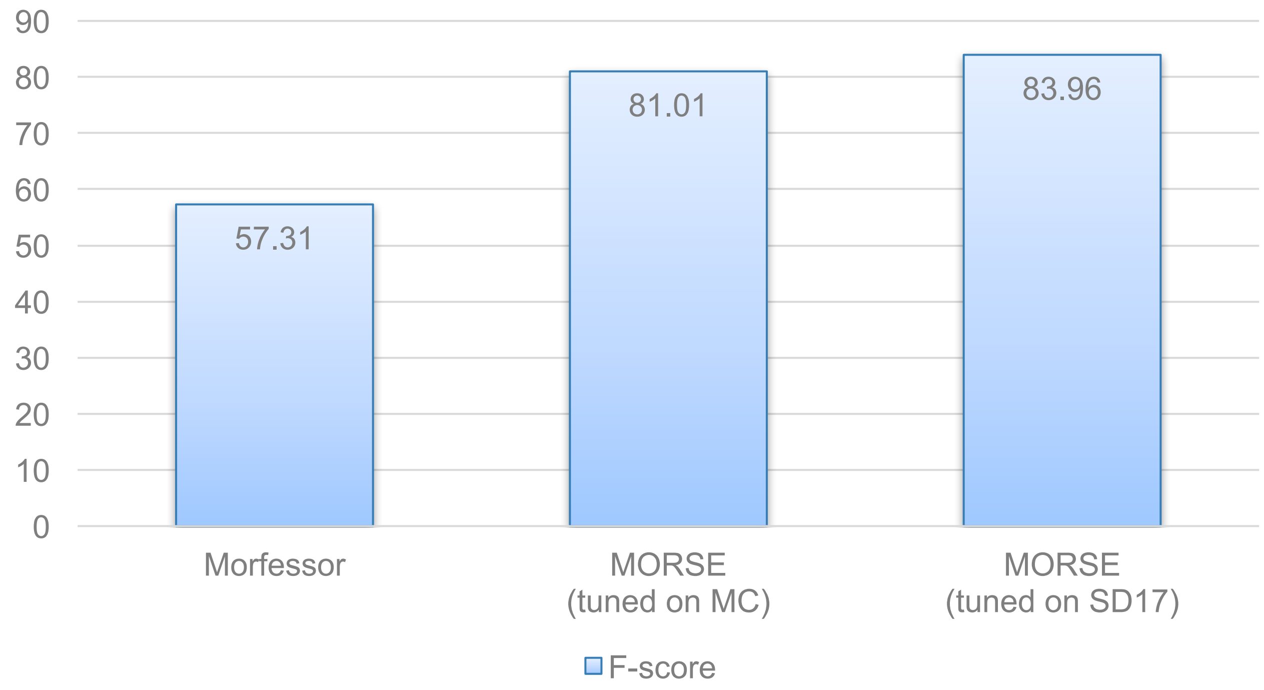
Non-compositional	Business	
Trivial instances	Turning-point	Player’s
Human error	Turning	

Experiments

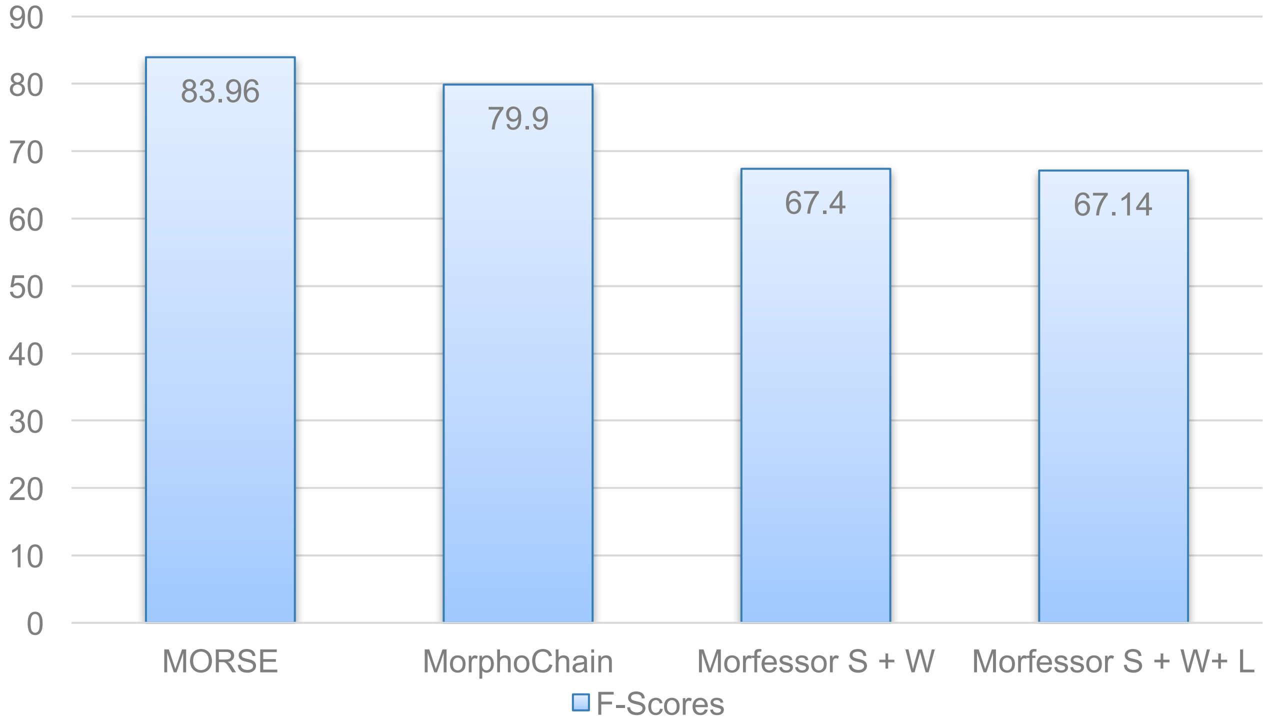
New Dataset: SD17

- 2000 words
- Compositional
- 91% inter-annotator agreement
- In canonical (butterfly + ies) and non-canonical version (butterfl + ies)

Results on SD17

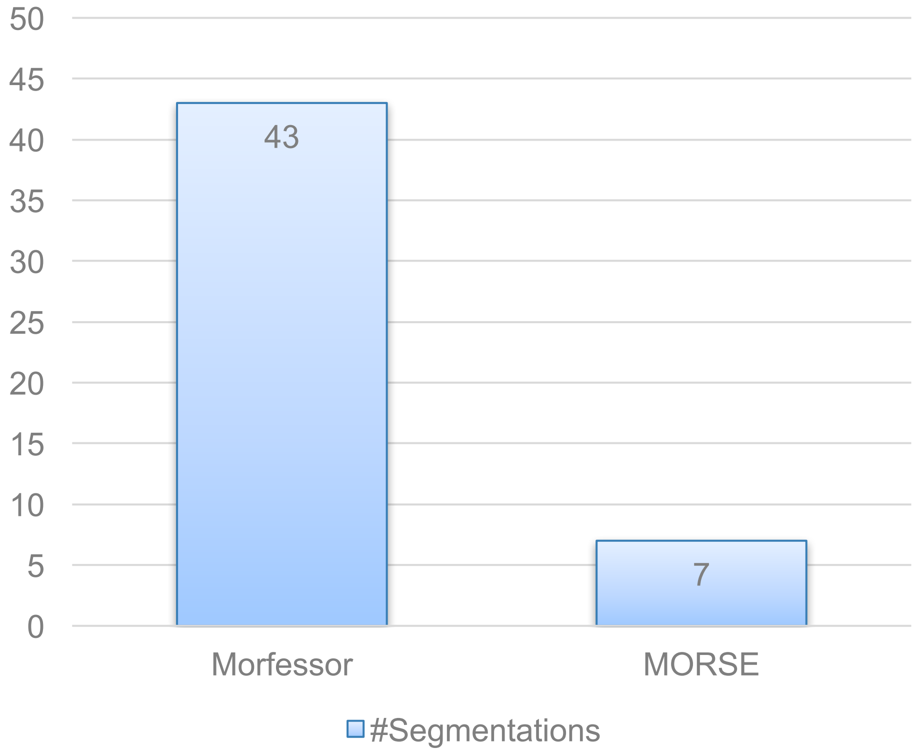


Against state-of-the-art



Negative Dataset

- 100 words like: honeymoon, passport, outdoors
- Checks for robustness



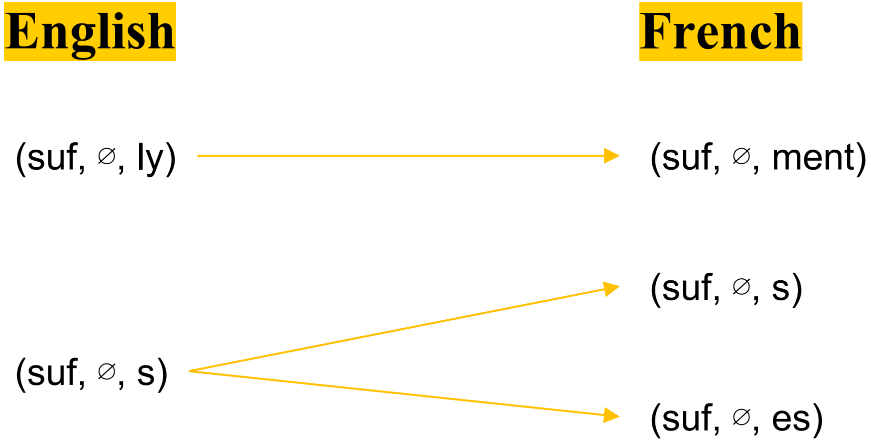
Looking forward

- Robustness to highly agglutinative languages
- Extending to other languages (non-concatenative)

k ai **t** aA **b** a

Looking forward

- Morphological mappings across languages



English

French

(suf, ∅, ly) → (suf, ∅, ment)

(suf, ∅, s) → (suf, ∅, s)

(suf, ∅, s) → (suf, ∅, es)

Links

IBM Watson

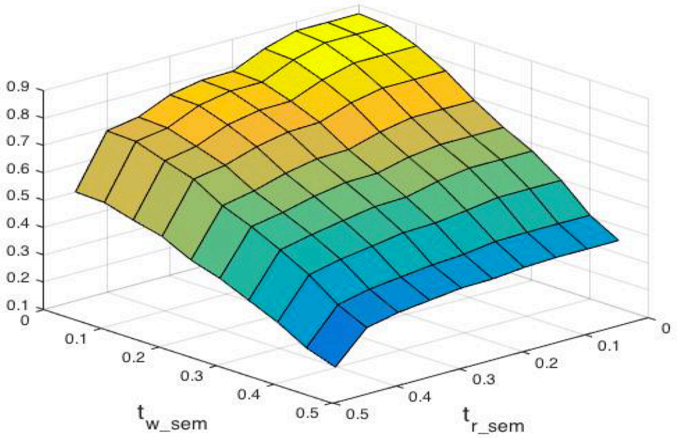
https://morse.mybluemix.net

https://github.com/yoonlee95/morse_segmentation

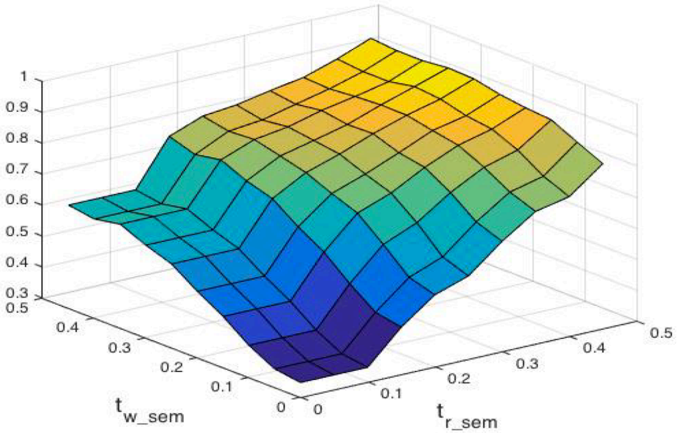
Thank you

Questions?

Effect of Hyperparameters



Recall



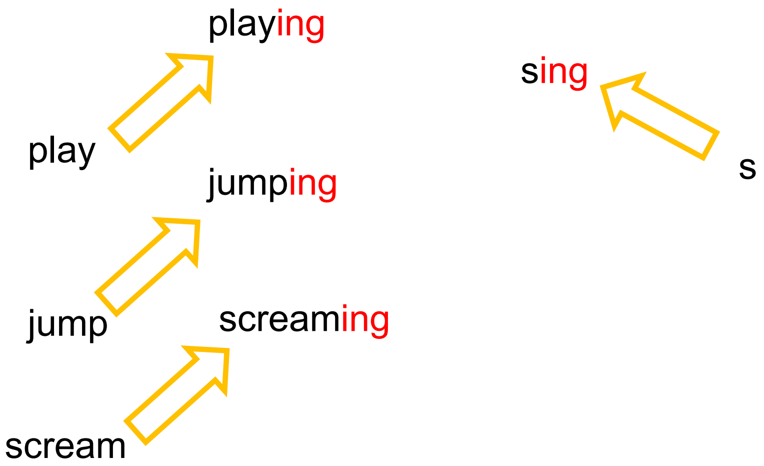
Precision

Prerequisite

Morpho-syntactic regularities in word vectors

Valid rule with an **invalid instance**

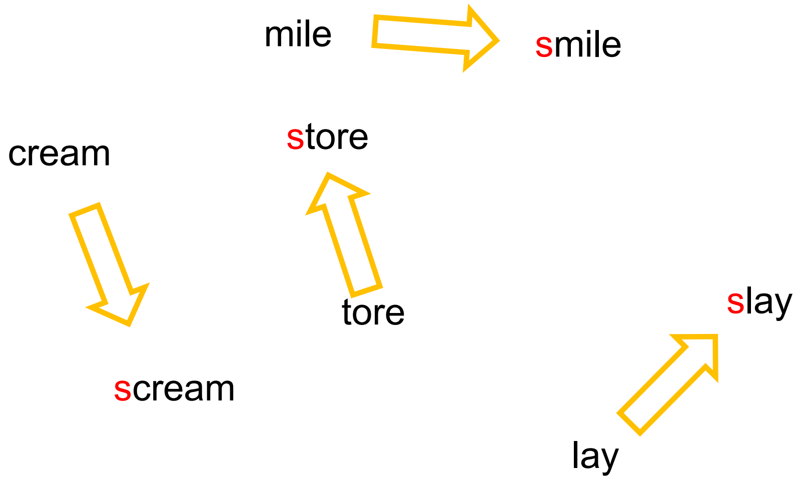
(suf, ∅, ing) **(s, sing)**



playing
sing
play
s
jumping
jump
screaming
scream

Invalid rule

(pre, ∅, s)



mile
smile
cream
store
tore
scream
slay
lay

4

Demo

morse.mybluemix.net