May All Your Wishes Come True: A Study of Wishes and How to Recognize Them

Andrew B. Goldberg, Nathanael Fillmore, David Andrzejewski,
Zhiting Xu, Bryan Gibson & Xiaojin Zhu

Computer Sciences Department
University of Wisconsin-Madison

Times Square
Virtual Wishing Well

- In December 2007, Web users sent in their wishes for the new year
- Wishes were printed on confetti
- Released from the sky at midnight in sync with the famous “ball drop”
- Over 100,000 wishes collected to form the WISH corpus

Sample New Year's Wishes

Freq.	Wish
514	peace on earth
351	peace
331	world peace
244	happy new year
112	love
76	health and happiness
75	to be happy
51	i wish for world peace
21	i wish for health and happiness
21	let there be peace on earth
16	to find my true love

Freq.	Wish
8	i wish for a puppy
7	for the war in iraq to end
6	peace on earth please
5	a free democratic venezuela
5	may the best of 2007 be the worst of 2008
5	to be financially stable
1	a little goodness for everyone would be nice
1	i hope i get accepted into a college that i like
1	i wish to get more sex in 2008
1	please let name be healthy and live all year
1	to be emotionally stable and happy

What is a wish?

What is a wish?

Formally:

wish (n.) “a desire or hope for something to happen”

What is a wish?

Formally:

wish (n.) “a desire or hope for something to happen”

- Open questions in NLP:
  - How are wishes expressed?
  - How can wishful expressions be automatically recognized?

What is a wish?

Formally:

wish (n.) “a desire or hope for something to happen”

- Open questions in NLP:
  - How are wishes expressed?
  - How can wishful expressions be automatically recognized?
- Our work:
  - Analyze this unique new collection of wishes
  - Leverage the WISH corpus to build general “wish detectors”
  - Demonstrate effectiveness on consumer product reviews and informal political discussion online

Outline

Outline

- Why study wishes? (relation to prior work)
  - Sentiment analysis
  - Psychology / cognitive science

Outline

- Why study wishes? (relation to prior work)
  - Sentiment analysis
  - Psychology / cognitive science
- Analysis of the WISH corpus
  - Topic and scope of wishes
  - Geographical differences
  - Latent topic modeling

Outline

- Why study wishes? (relation to prior work)
  - Sentiment analysis
  - Psychology / cognitive science
- Analysis of the WISH corpus
  - Topic and scope of wishes
  - Geographical differences
  - Latent topic modeling
- Building wish detectors
  - Key contribution: Automatically discovering wish templates

Outline

- Why study wishes? (relation to prior work)
  - Sentiment analysis
  - Psychology / cognitive science
- Analysis of the WISH corpus
  - Topic and scope of wishes
  - Geographical differences
  - Latent topic modeling
- Building wish detectors
  - Key contribution: Automatically discovering wish templates
- Experimental results

Why study wishes?

Why study wishes?

- Wishes add a novel dimension to sentiment analysis, opinion mining
  - What people explicitly want, not just what they like or dislike

> "Great camera. Indoor shots with a flash are not quite as good as 35mm. I wish the camera had a higher optical zoom so that I could take even better wildlife photos."

  - Automatic "wish detector" can provide political value & business intelligence

Why study wishes?

- Wishes add a novel dimension to sentiment analysis, opinion mining
  - What people explicitly want, not just what they like or dislike

> "Great camera. Indoor shots with a flash are not quite as good as 35mm. I wish the camera had a higher optical zoom so that I could take even better wildlife photos."

  - Automatic "wish detector" can provide political value & business intelligence

- Wishes can reveal a lot about people
  - Psychologists have studied wish content vs. location, gender, age, etc (Speer 1939, Milgram and Riedel 1969, Ehrlichman and Eichenstein 1992, King and Broyles 1997)
  - WISH corpus: much larger scale, from the entire globe

The WISH corpus

Analysis of the WISH corpus

Analysis of the WISH corpus

- Almost 100,000 wishes collected over 10 days in December 2007
  - We focus on the 89,574 wishes written in English
  - Remaining 10,000+ in Portuguese, Spanish, Chinese, French, etc

Analysis of the WISH corpus

- Almost 100,000 wishes collected over 10 days in December 2007
  - We focus on the 89,574 wishes written in English
  - Remaining 10,000+ in Portuguese, Spanish, Chinese, French, etc
- Many contain optional state/country location entered by the wisher

Analysis of the WISH corpus

- Almost 100,000 wishes collected over 10 days in December 2007
  - We focus on the 89,574 wishes written in English
  - Remaining 10,000+ in Portuguese, Spanish, Chinese, French, etc
- Many contain optional state/country location entered by the wisher
- Minimal preprocessing
  - TreeBank tokenization, downcasing, punctuation removal

Analysis of the WISH corpus

- Almost 100,000 wishes collected over 10 days in December 2007
  - We focus on the 89,574 wishes written in English
  - Remaining 10,000+ in Portuguese, Spanish, Chinese, French, etc
- Many contain optional state/country location entered by the wisher
- Minimal preprocessing
  - TreeBank tokenization, downcasing, punctuation removal
- Each wish is treated as a single entity (even if multiple sentences)

Analysis of the WISH corpus

- Almost 100,000 wishes collected over 10 days in December 2007
  - We focus on the 89,574 wishes written in English
  - Remaining 10,000+ in Portuguese, Spanish, Chinese, French, etc
- Many contain optional state/country location entered by the wisher
- Minimal preprocessing
  - TreeBank tokenization, downcasing, punctuation removal
- Each wish is treated as a single entity (even if multiple sentences)
- Average length of wishes is 8 tokens

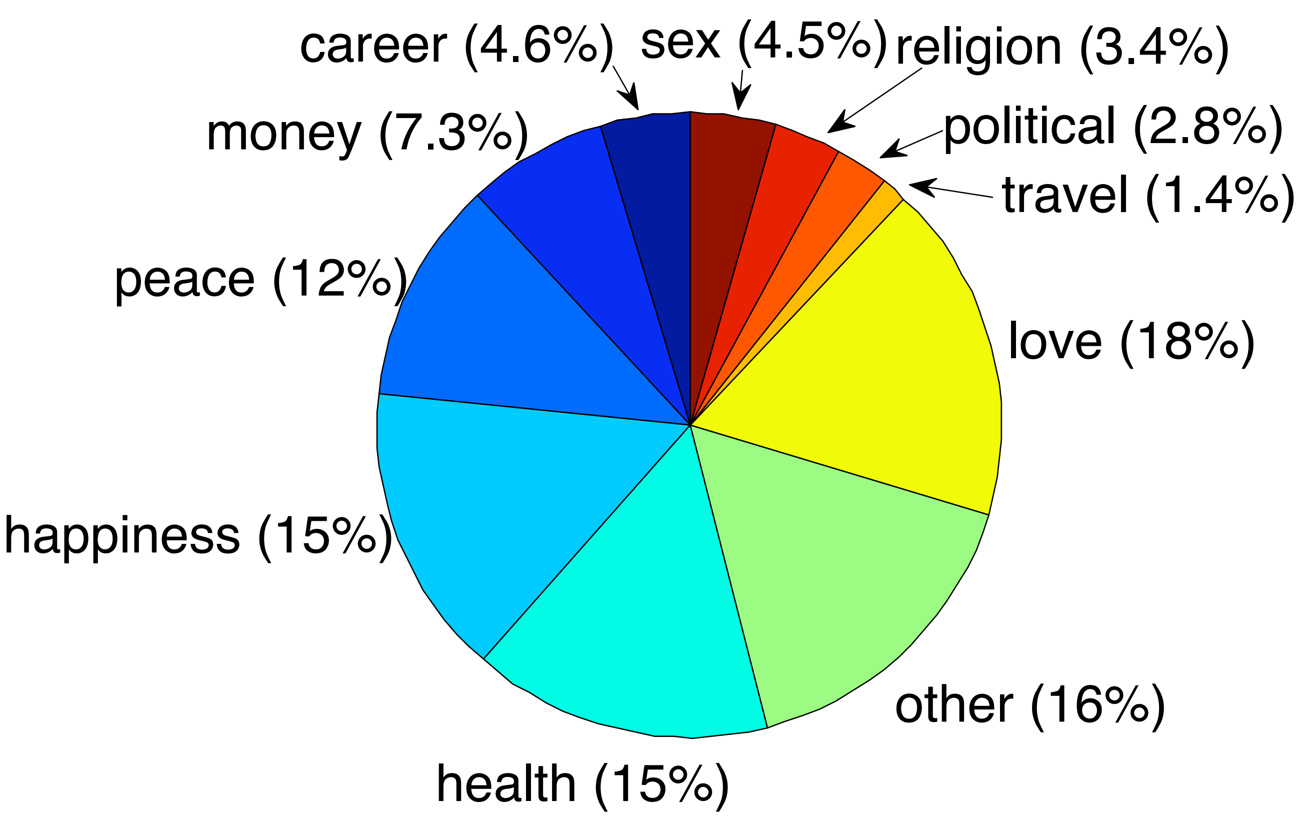
WISH corpus: Scope and topic of wishes

Manually annotated random subsample of 5,000 wishes

WISH corpus: Scope and topic of wishes

Manually annotated random subsample of 5,000 wishes

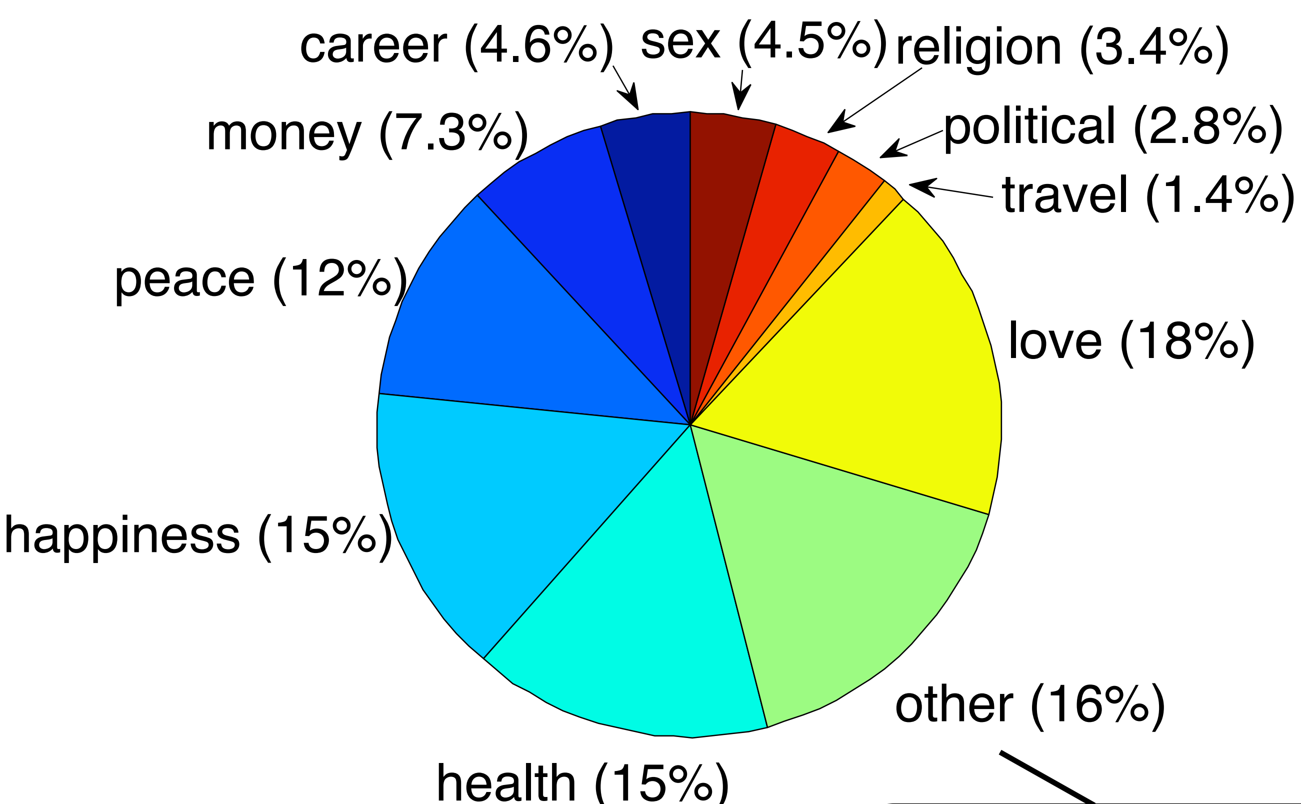
Topic of wishes:
what the wish is about



WISH corpus: Scope and topic of wishes

Manually annotated random subsample of 5,000 wishes

Topic of wishes:
what the wish is about

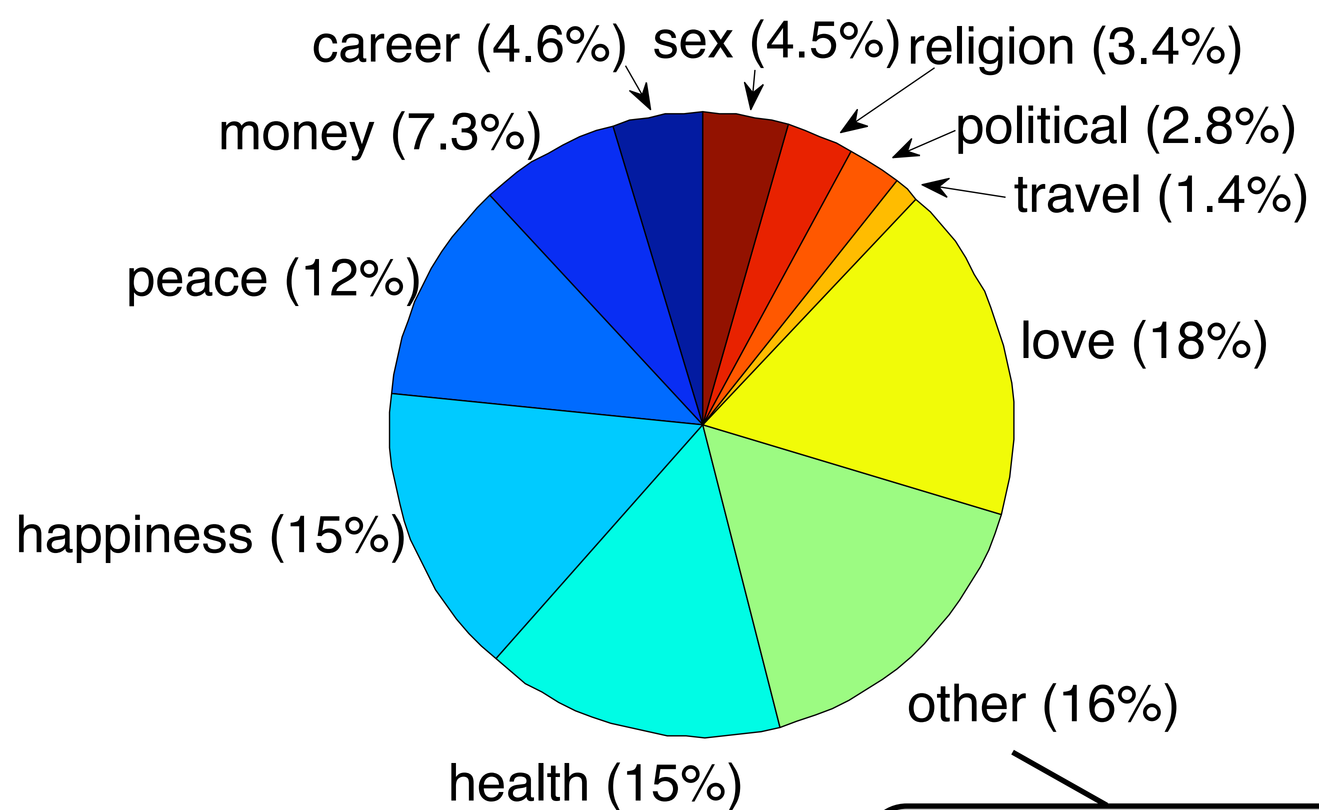


individual requests: “i wish for a new puppy”
solicitations: “call me *555-1234*”, “visit *website*.com”
sinister: “to take over the world”

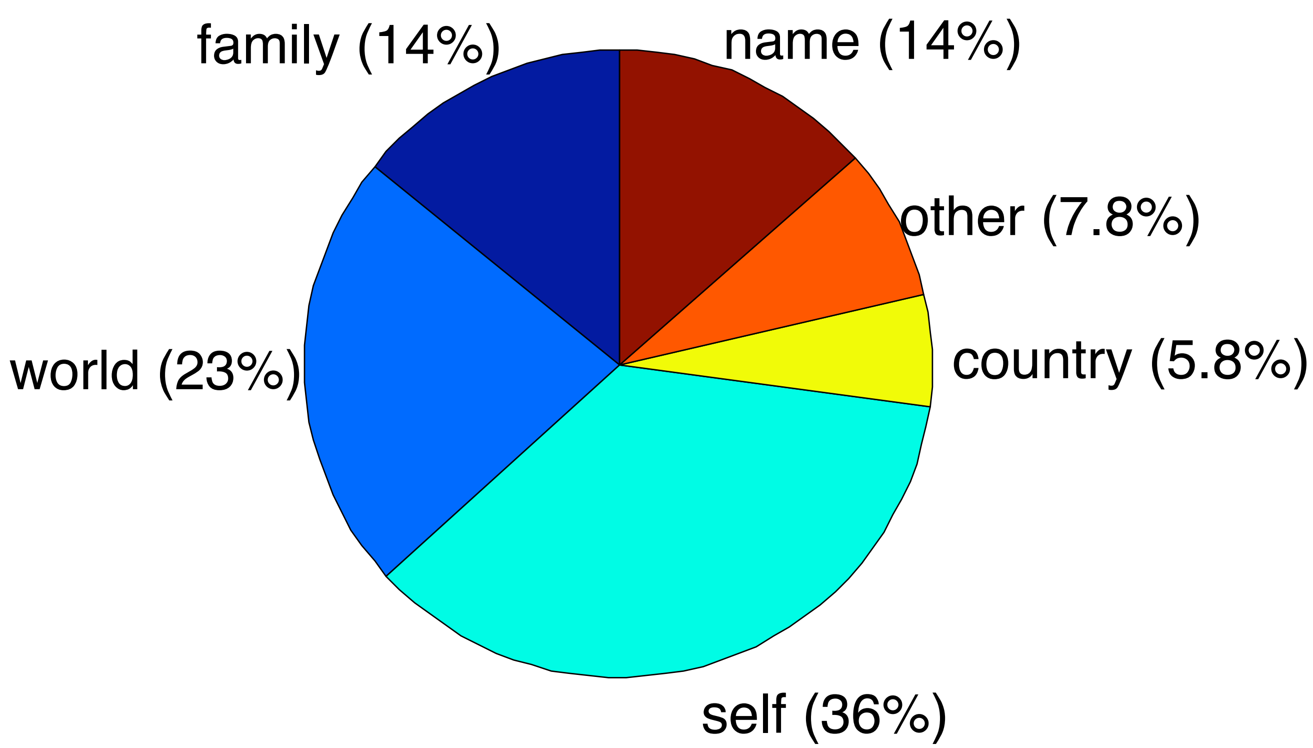
WISH corpus: Scope and topic of wishes

Manually annotated random subsample of 5,000 wishes

Topic of wishes:
what the wish is about



Scope of wishes:
who the wish is aimed at



individual requests: “i wish for a new puppy”
solicitations: “call me *555-1234*”, “visit *website*.com”
sinister: “to take over the world”

WISH corpus: Geographical differences

WISH corpus: Geographical differences

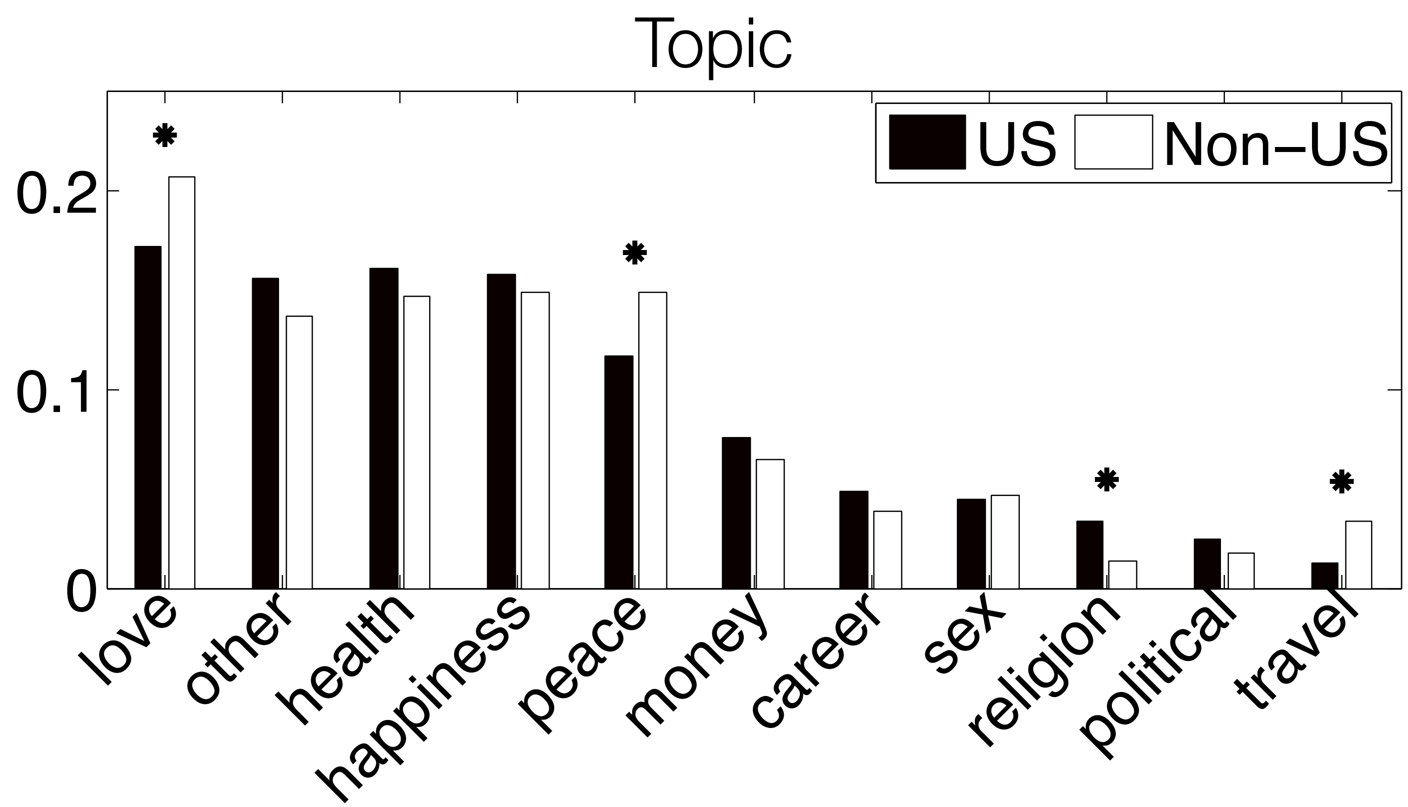
- About 4,000 of the manually annotated wishes included valid location information
  - Covered all 50 U.S. states and all continents except Antarctica

WISH corpus: Geographical differences

- About 4,000 of the manually annotated wishes included valid location information
  - Covered all 50 U.S. states and all continents except Antarctica
- We compared topic and scope distributions between U.S. and non-U.S. wishes

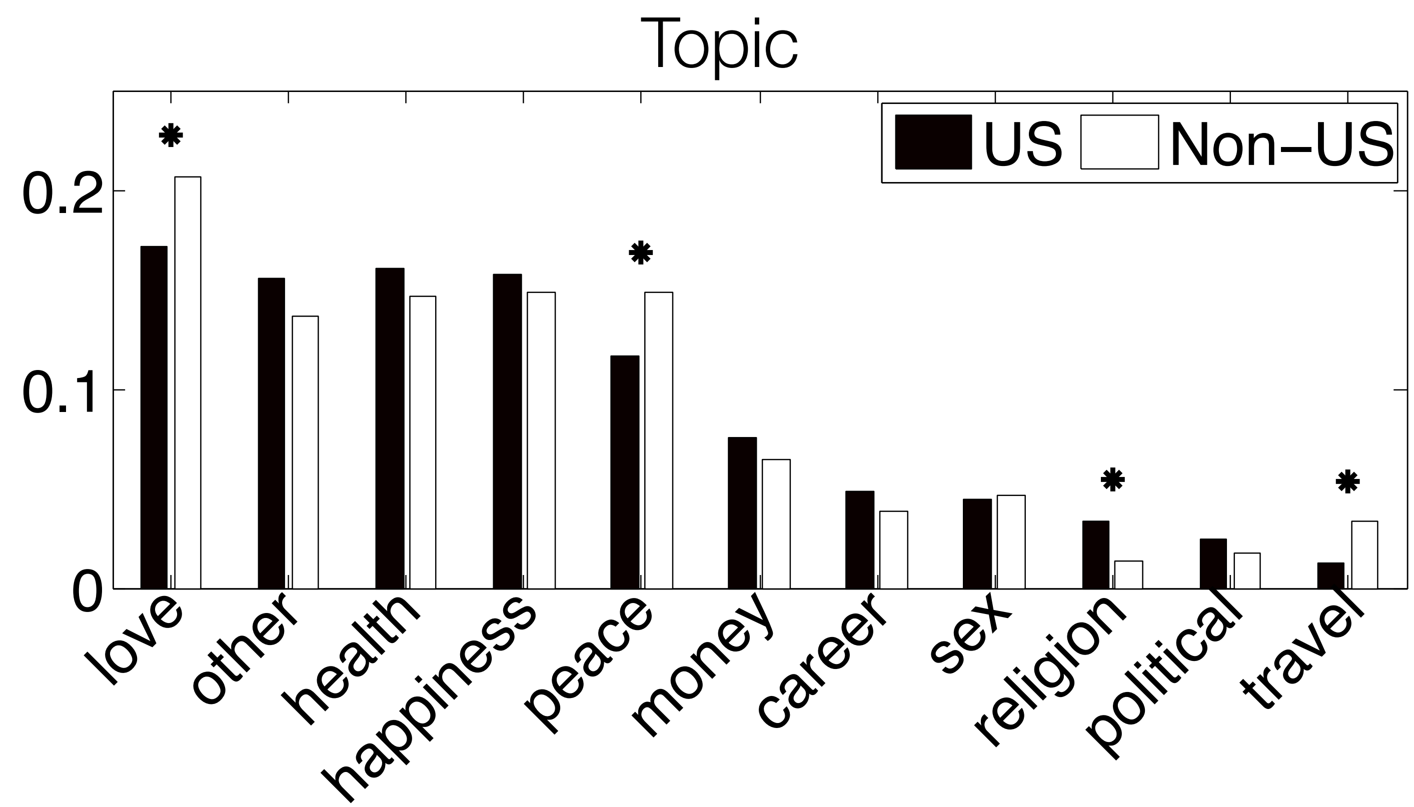
WISH corpus: Geographical differences

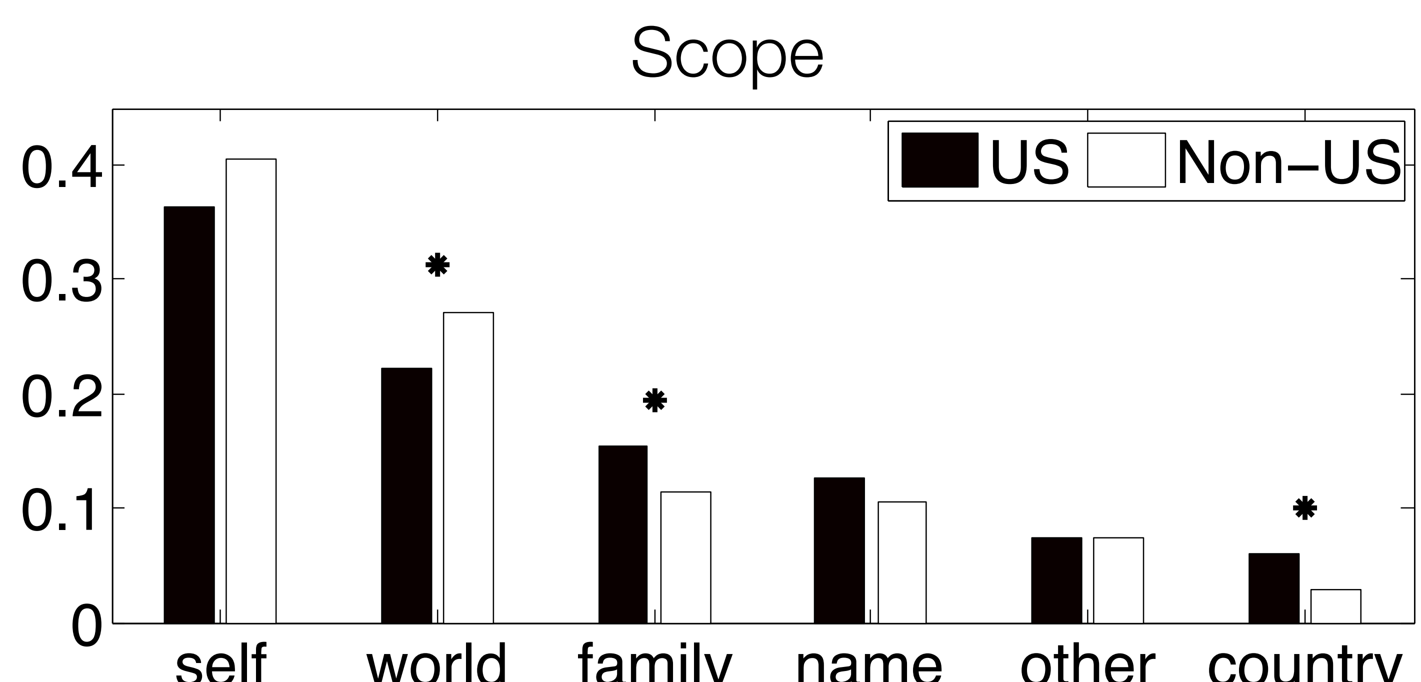
- About 4,000 of the manually annotated wishes included valid location information
  - Covered all 50 U.S. states and all continents except Antarctica
- We compared topic and scope distributions between U.S. and non-U.S. wishes



WISH corpus: Geographical differences

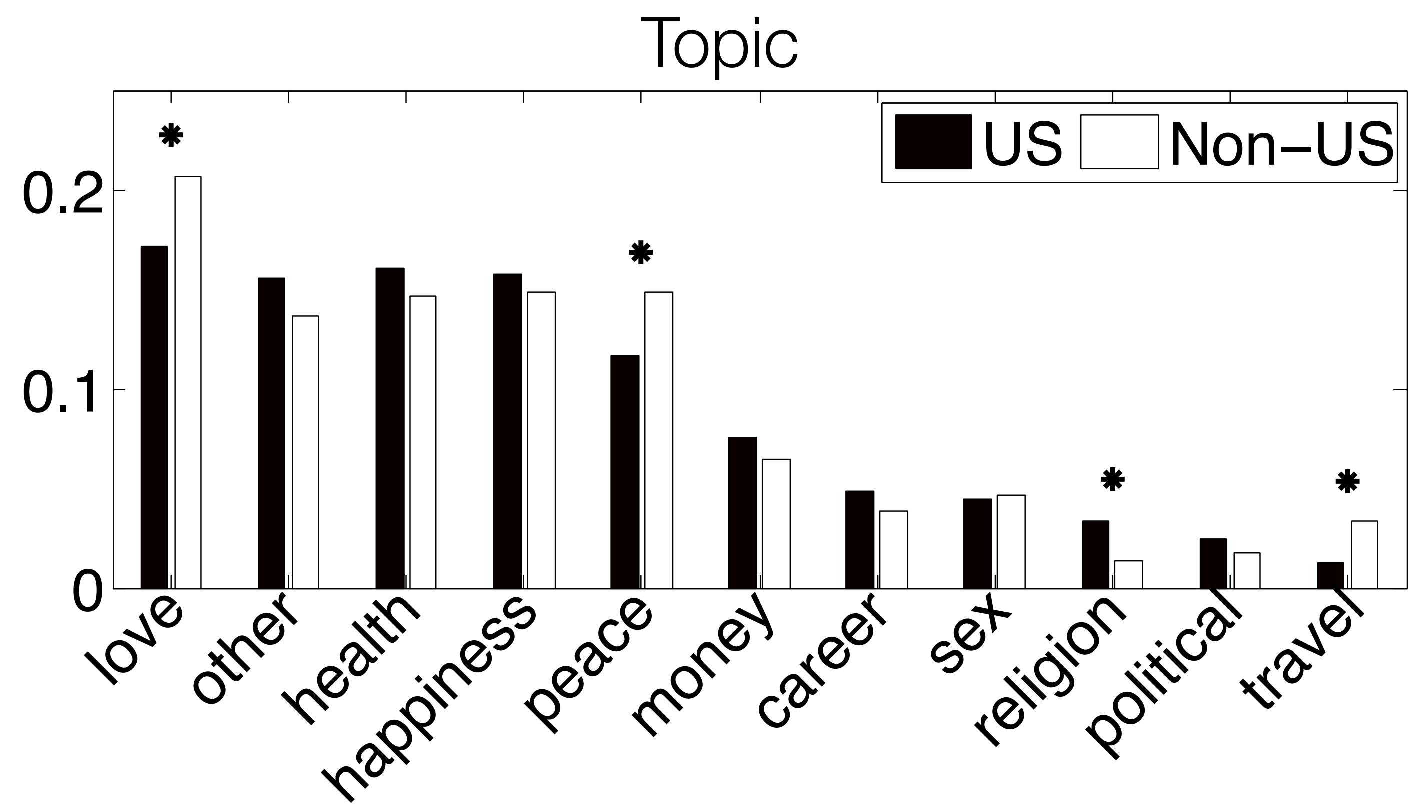
- About 4,000 of the manually annotated wishes included valid location information
  - Covered all 50 U.S. states and all continents except Antarctica
- We compared topic and scope distributions between U.S. and non-U.S. wishes

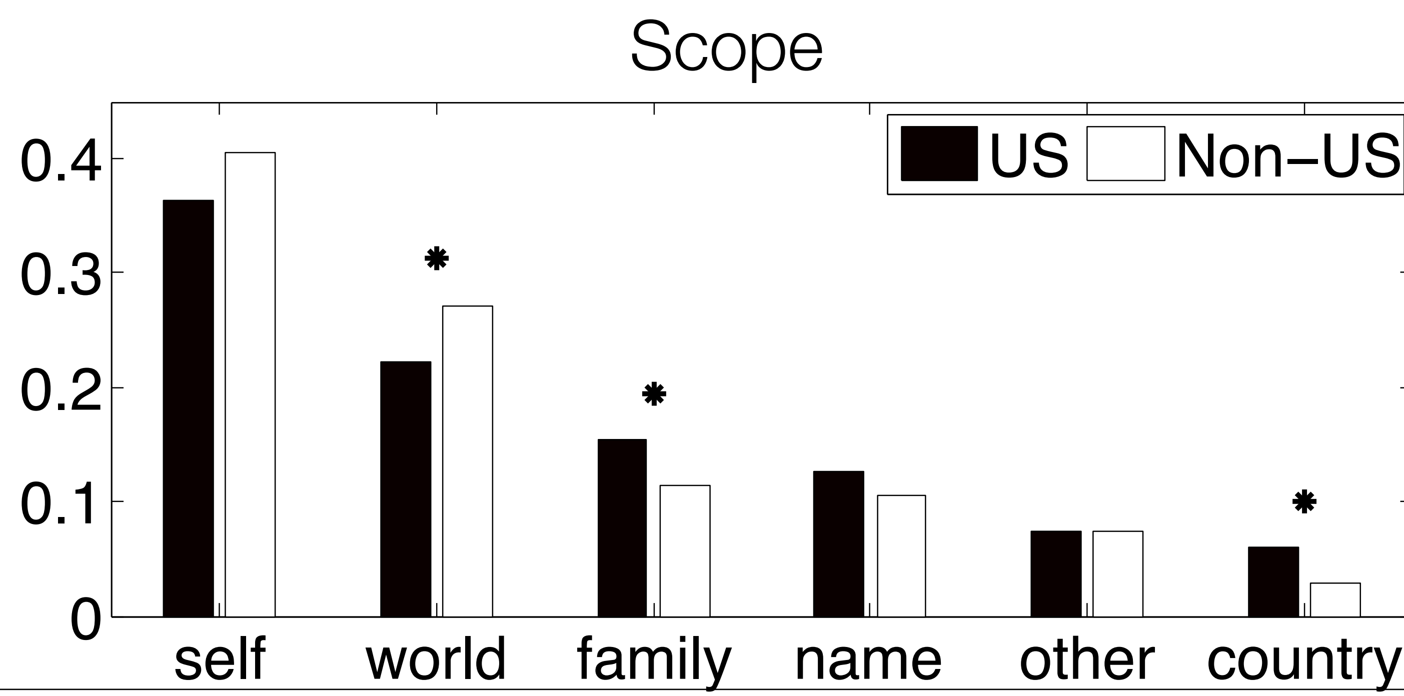




WISH corpus: Geographical differences

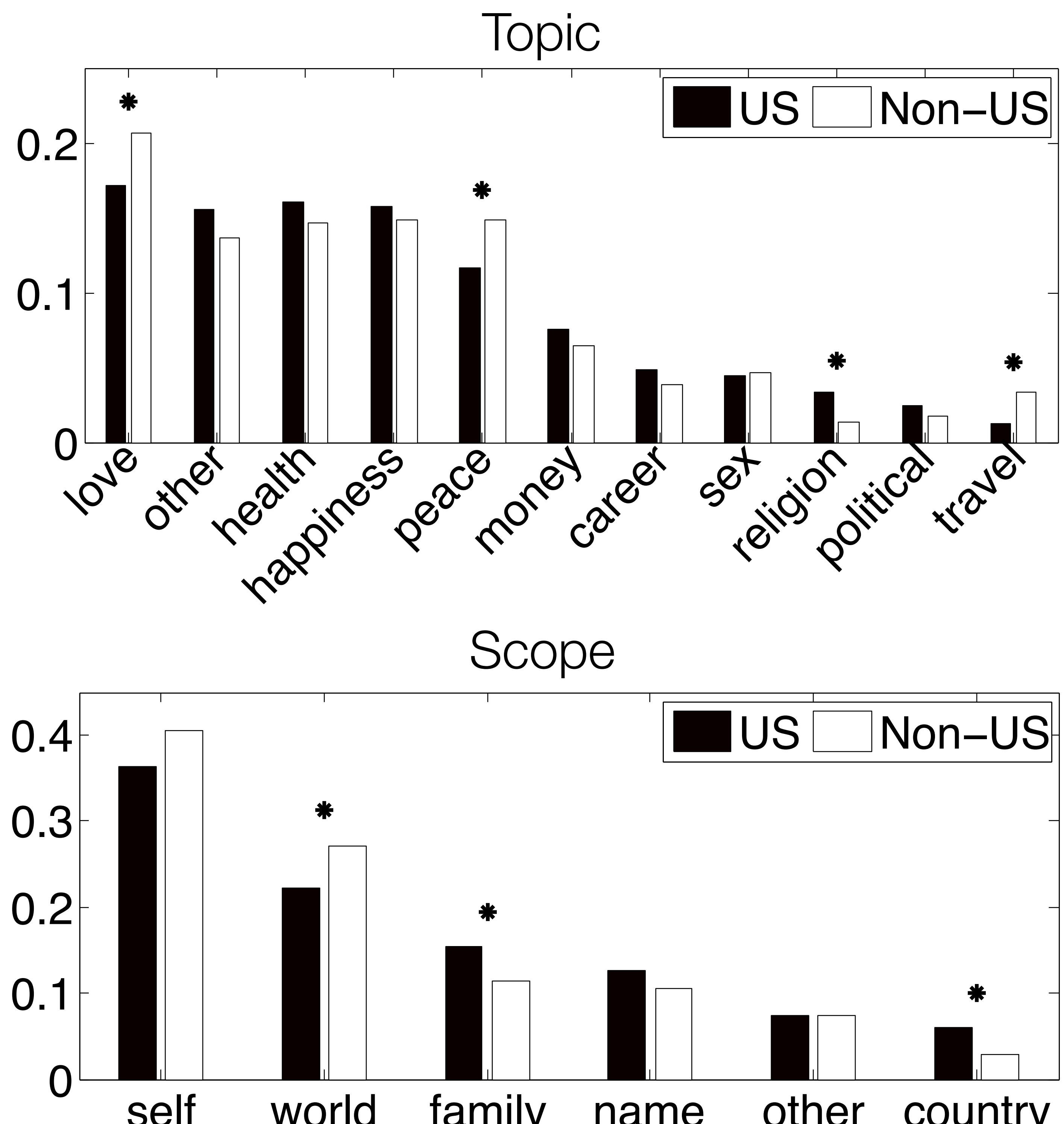
- About 4,000 of the manually annotated wishes included valid location information
  - Covered all 50 U.S. states and all continents except Antarctica
- We compared topic and scope distributions between U.S. and non-U.S. wishes
- Statistically significant differences in both cases (Pearson X^2-test, $p < 0.01$)





WISH corpus: Geographical differences

- About 4,000 of the manually annotated wishes included valid location information
  - Covered all 50 U.S. states and all continents except Antarctica
- We compared topic and scope distributions between U.S. and non-U.S. wishes
- Statistically significant differences in both cases (Pearson X^2-test, $p < 0.01$)
- *But* no significant difference between red vs. blue states



WISH corpus: Latent topic modeling

WISH corpus: Latent topic modeling

- So far analysis was of 5,000 manually labeled wishes

WISH corpus: Latent topic modeling

- So far analysis was of 5,000 manually labeled wishes
- We automatically analyzed all ~90,000 using Latent Dirichlet Allocation
  - Each wish is treated as a short document
  - 12 topics
  - Inference performed by collapsed Gibbs sampling
  - Hyperparameters set to α=0.5, β=0.1

WISH corpus: Latent topic modeling

Topic	Top words, sorted by p(word|topic)	Subjective Label
1	year, new, happy, 2008, best, everyone, great, wishing, hope	New Year
2	all, god, home, come, safe, us, bless, troops, bring, iraq, return	Troops
3	end, no, more, 2008, war, president, paul, ron, less, bush, vote	Election
4	more, better, life, one, live, time, make, people, than, day, every	Life
5	health, happiness, good, family, friends, prosperity, wealth, success	Prosperity
6	love, find, true, life, meet, want, man, marry, someone, boyfriend	Love
7	get, job, out, hope, school, better, house, well, back, college	Career
8	win, 2008, money, want, make, become, lottery, more, great, lots	Money
9	peace, world, love, earth, happiness, everyone, joy, 2008, around	Peace
10	love, forever, jesus, know, together, u, always, best, mom, christ	Religion
11	healthy, family, baby, life, children, safe, husband, stay, marriage	Family
12	me, lose, please, let, cancer, weight, cure, mom, mother, visit, dad	Health

Building wish detectors

Building wish detectors

Novel NLP task: Wish Detection

Given sentence *S*, classify *S* as wish or non-wish

Building wish detectors

Novel NLP task: Wish Detection
Given sentence *S*, classify *S* as wish or non-wish

- Want an approach that will extend beyond New Year's wishes
    - Target domains: product reviews, political discussions

Building wish detectors

Novel NLP task: Wish Detection
Given sentence *S*, classify *S* as wish or non-wish

- Want an approach that will extend beyond New Year's wishes
  - Target domains: product reviews, political discussions
- Wishes are highly domain dependent
  - New Year's eve: "I wish for world peace"
  - Product review: "I want to have instant access to the volume"

Building wish detectors

Novel NLP task: Wish Detection
Given sentence *S*, classify *S* as wish or non-wish

- Want an approach that will extend beyond New Year's wishes
  - Target domains: product reviews, political discussions
- Wishes are highly domain dependent
  - New Year's eve: "I wish for world peace"
  - Product review: "I want to have instant access to the volume"
- Initial study
  - Assume some labeled data in target domains
  - Try to beat some standard baselines by exploiting the WISH corpus to learn patterns of wish expressions (wish templates)

Two simple baseline wish detectors

- Do not use WISH corpus

Two simple baseline wish detectors

- Do not use WISH corpus

Manual

- Rule-based classifier
- If part of a sentence matches a template, classify it as a wish
- Some of the 13 templates created by two native English speakers:

i wish __	if only __
i hope __	would be better if __
i want __	would like if __
hopefully __	should __

Expect high precision, low recall

Two simple baseline wish detectors

- Do not use WISH corpus

Manual

- Rule-based classifier
- If part of a sentence matches a template, classify it as a wish
- Some of the 13 templates created by two native English speakers:

i wish __	if only __
i hope __	would be better if __
i want __	would like if __
hopefully __	should __

Expect high precision, low recall

Words

- Linear Support Vector Machine
- Train on labeled training set from the target domain
- Representation:
  - binary word-indicator vector
  - normalized to sum to 1
- Natural first baseline for a new text classification task

Expect high recall, low precision

Learning wish templates

Key idea: Exploit redundancy in how wishes are expressed

Learning wish templates

Key idea: Exploit redundancy in how wishes are expressed

Many entries in the WISH corpus contain only a short “wish content”

world peace   health and happiness

Learning wish templates

Key idea: Exploit redundancy in how wishes are expressed

Many entries in the WISH corpus contain only a short “wish content”

world peace   health and happiness

These “wish contents” appear within longer wishes with a common prefix/suffix:

i wish for world peace   **i wish for** health and happiness

Learning wish templates

Key idea: Exploit redundancy in how wishes are expressed

Many entries in the WISH corpus contain only a short “wish content”

world peace health and happiness

These “wish contents” appear within longer wishes with a common prefix/suffix:

i wish for world peace **i wish for** health and happiness

Intuitively, popular content appears within popular templates.

Learning wish templates

Key idea: Exploit redundancy in how wishes are expressed

Many entries in the WISH corpus contain only a short “wish content”

world peace   health and happiness

These “wish contents” appear within longer wishes with a common prefix/suffix:

i wish for world peace   **i wish for** health and happiness

Intuitively, popular content appears within popular templates.

Can discover non-obvious templates, too:

world peace, peace on earth → **let there be ___**

become rich, win the lottery → **to finally ___**

get a job, save the environment → **___ please**

Learning wish templates

Formally, we build a bipartite graph

Two kinds of nodes: Content nodes $c \in C$ on left, Template nodes $t \in T$ on right

Two kinds of edges:
- $c \rightarrow t$ (weighted by # times content appears in the template)
- $t \rightarrow c$ (weighted by # times template matches a content node)

Content

Templates

Learning wish templates

Formally, we build a bipartite graph

Two kinds of nodes: Content nodes $c \in C$ on left, Template nodes $t \in T$ on right

Two kinds of edges:

- $c \rightarrow t$ (weighted by # times content appears in the template)
- $t \rightarrow c$ (weighted by # times template matches a content node)

Learning wish templates

Formally, we build a bipartite graph

Two kinds of nodes: Content nodes $c \in C$ on left, Template nodes $t \in T$ on right

Two kinds of edges:
- $c \rightarrow t$ (weighted by # times content appears in the template)
- $t \rightarrow c$ (weighted by # times template matches a content node)

Learning wish templates

Formally, we build a bipartite graph

Two kinds of nodes: Content nodes $c \in C$ on left, Template nodes $t \in T$ on right

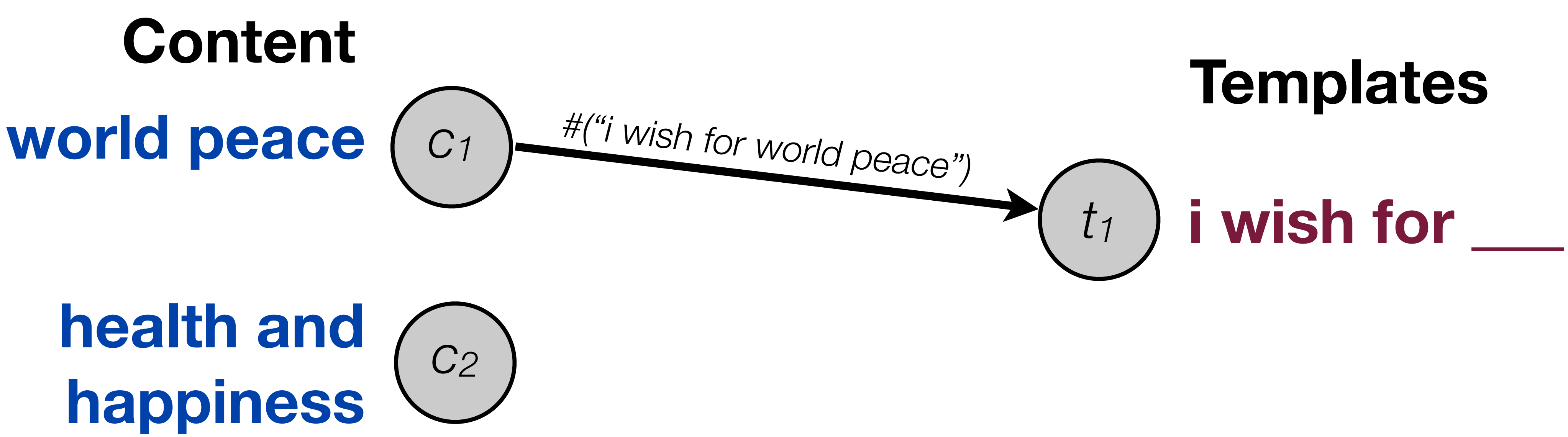
Two kinds of edges:
- $c \rightarrow t$ (weighted by # times content appears in the template)
- $t \rightarrow c$ (weighted by # times template matches a content node)

Learning wish templates

Formally, we build a bipartite graph

Two kinds of nodes: Content nodes $c \in C$ on left, Template nodes $t \in T$ on right

Two kinds of edges:
- $c \rightarrow t$ (weighted by # times content appears in the template)
- $t \rightarrow c$ (weighted by # times template matches a content node)

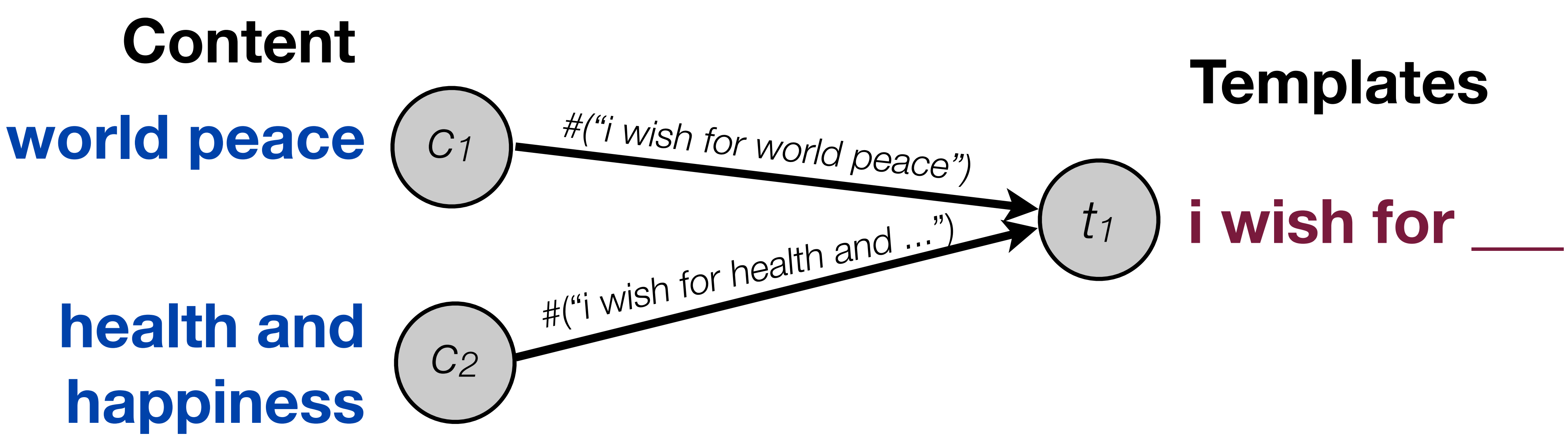


Learning wish templates

Formally, we build a bipartite graph

Two kinds of nodes: Content nodes $c \in C$ on left, Template nodes $t \in T$ on right

Two kinds of edges:
- $c \rightarrow t$ (weighted by # times content appears in the template)
- $t \rightarrow c$ (weighted by # times template matches a content node)

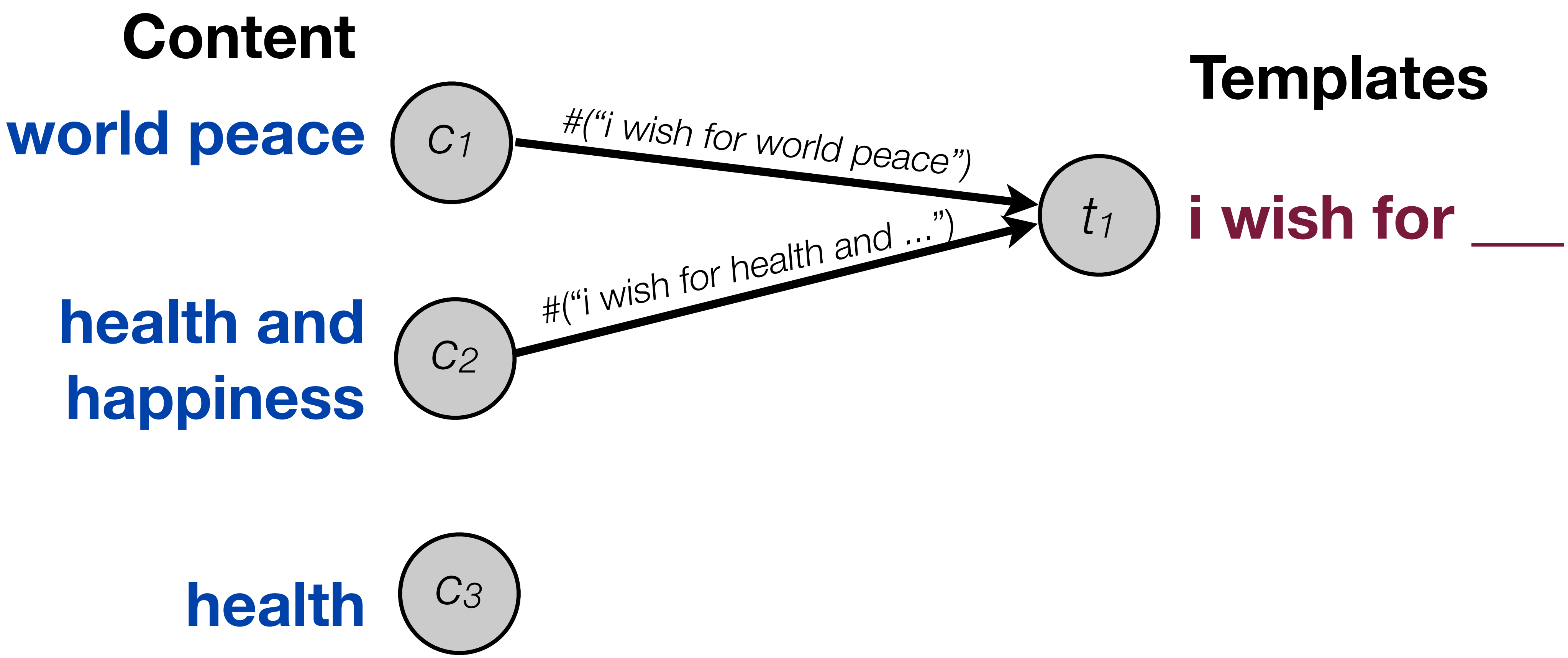


Learning wish templates

Formally, we build a bipartite graph

Two kinds of nodes: Content nodes $c \in C$ on left, Template nodes $t \in T$ on right

Two kinds of edges:
- $c \rightarrow t$ (weighted by # times content appears in the template)
- $t \rightarrow c$ (weighted by # times template matches a content node)

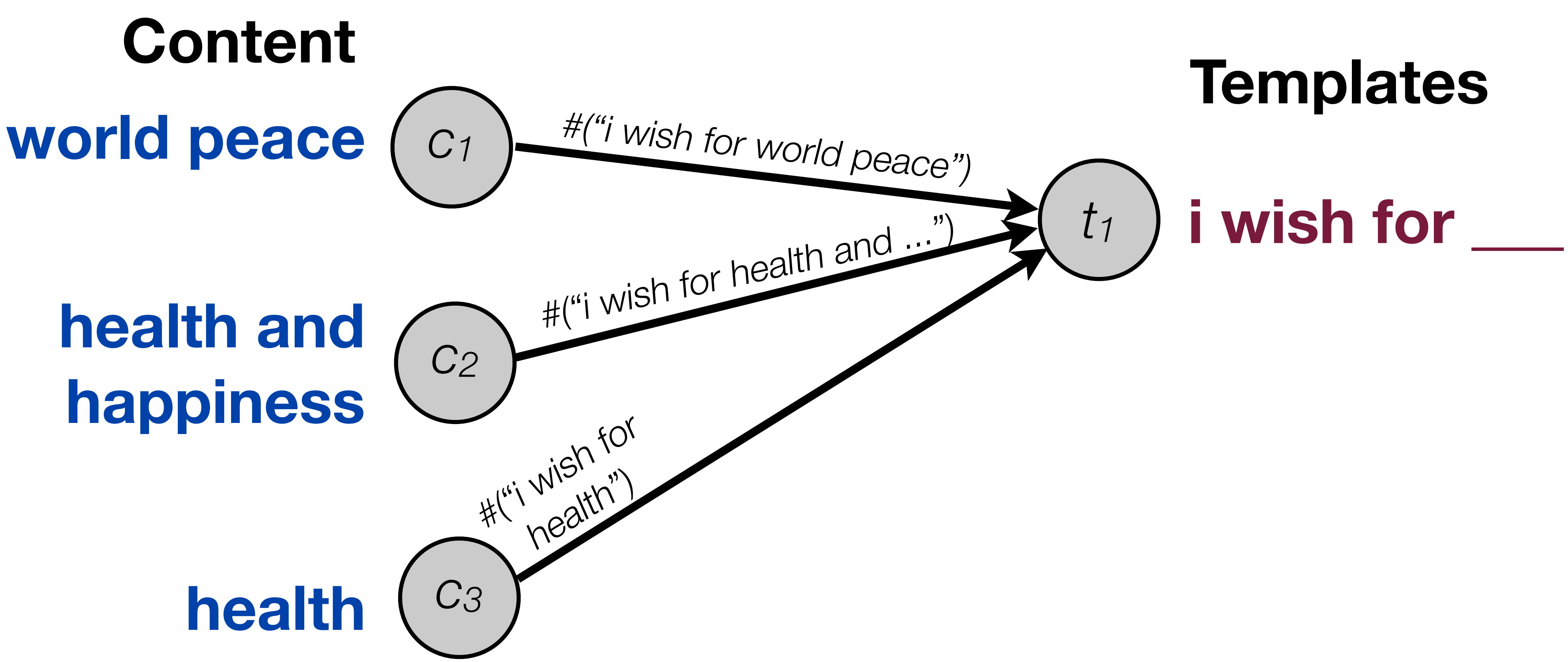


Learning wish templates

Formally, we build a bipartite graph

Two kinds of nodes: Content nodes $c \in C$ on left, Template nodes $t \in T$ on right

Two kinds of edges:
- $c \rightarrow t$ (weighted by # times content appears in the template)
- $t \rightarrow c$ (weighted by # times template matches a content node)

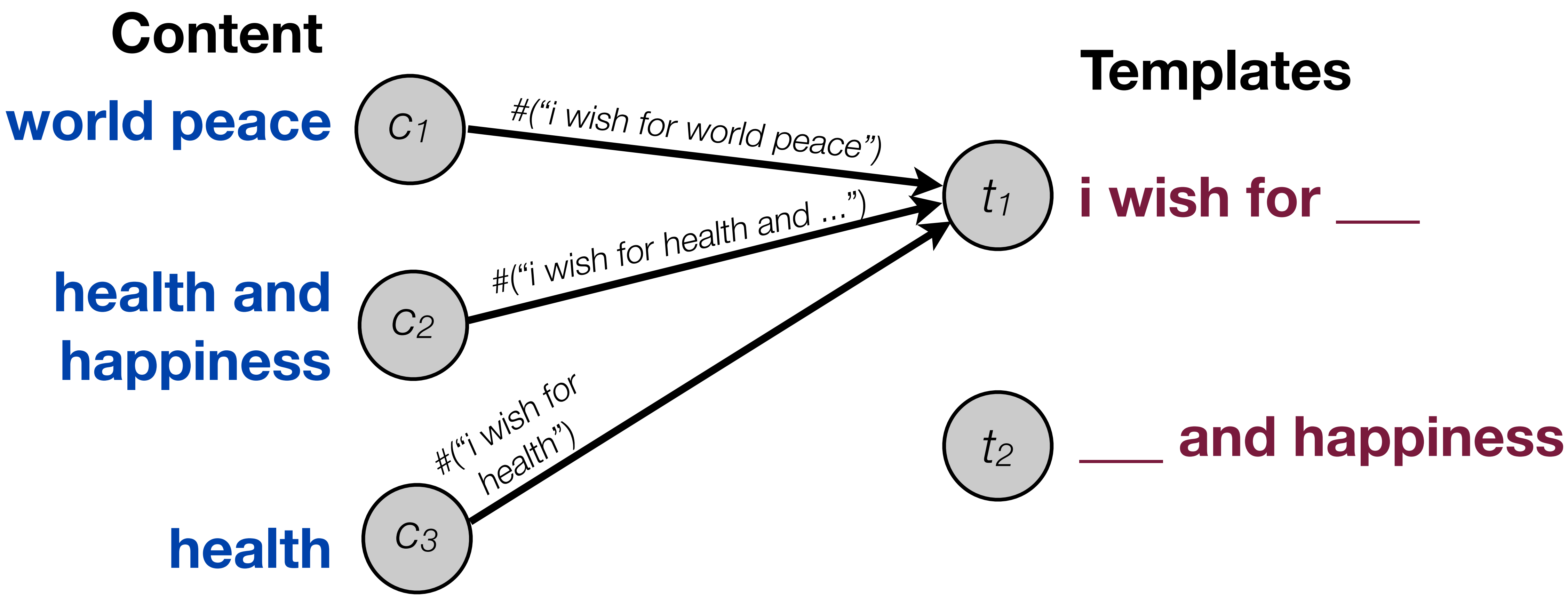


Learning wish templates

Formally, we build a bipartite graph

Two kinds of nodes: Content nodes $c \in C$ on left, Template nodes $t \in T$ on right

Two kinds of edges:
- $c \to t$ (weighted by # times content appears in the template)
- $t \to c$ (weighted by # times template matches a content node)

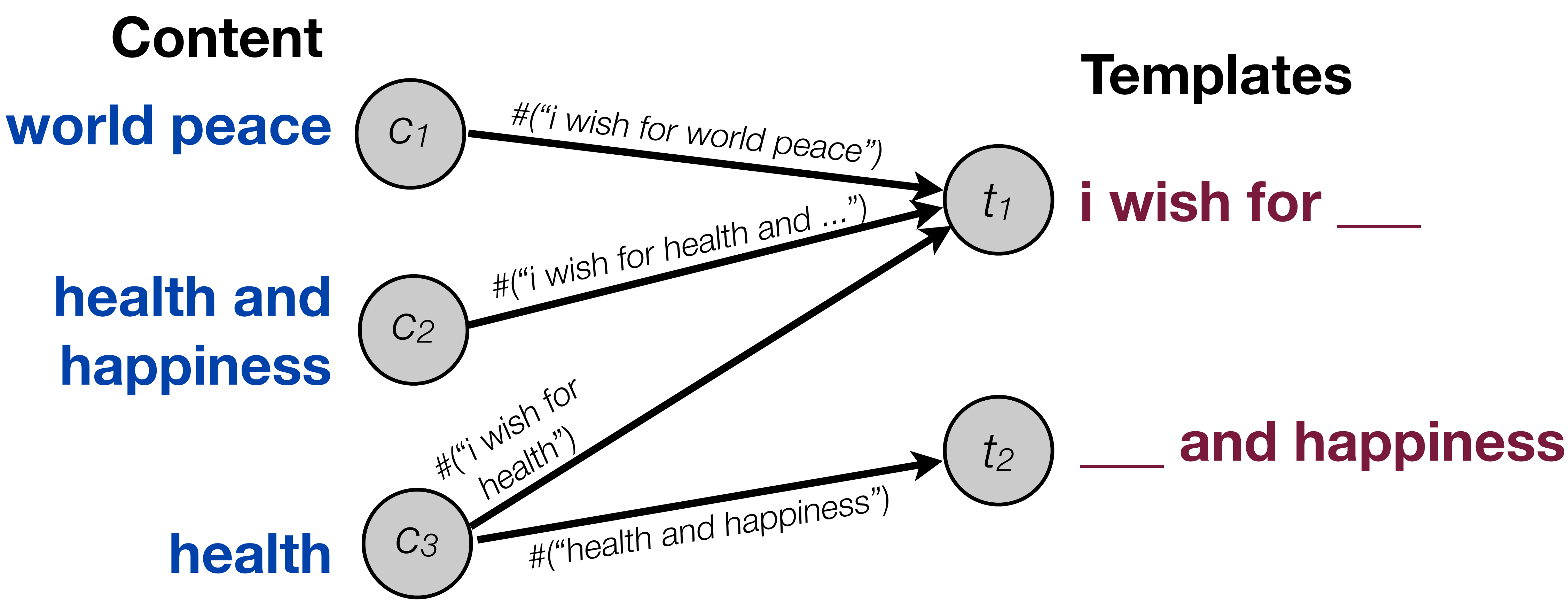


Learning wish templates

Formally, we build a bipartite graph

Two kinds of nodes: Content nodes $c \in C$ on left, Template nodes $t \in T$ on right

Two kinds of edges:
- $c \rightarrow t$ (weighted by # times content appears in the template)
- $t \rightarrow c$ (weighted by # times template matches a content node)

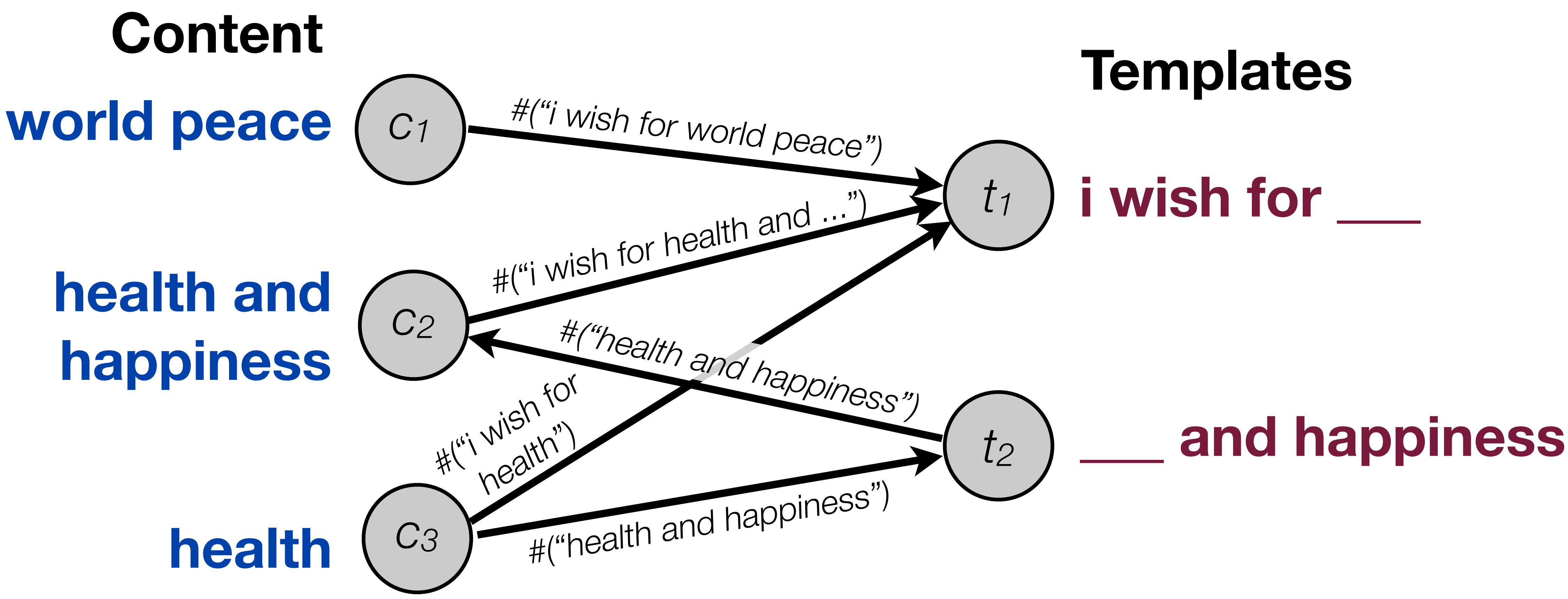


Learning wish templates

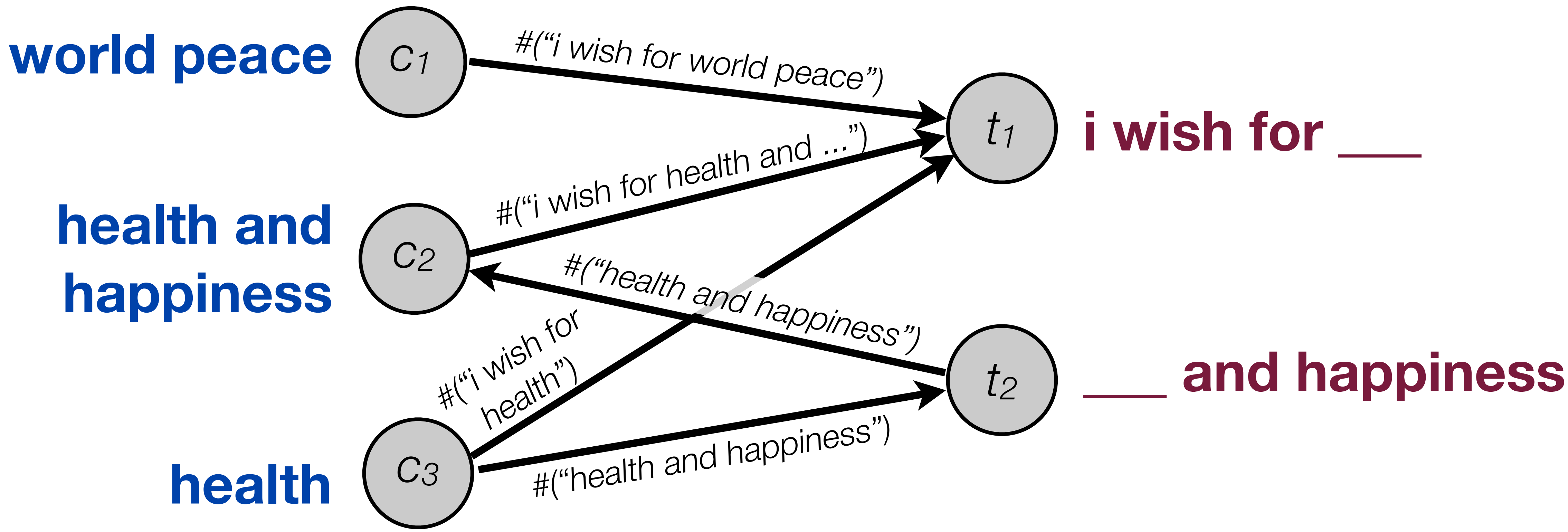
Formally, we build a bipartite graph

Two kinds of nodes: Content nodes $c \in C$ on left, Template nodes $t \in T$ on right

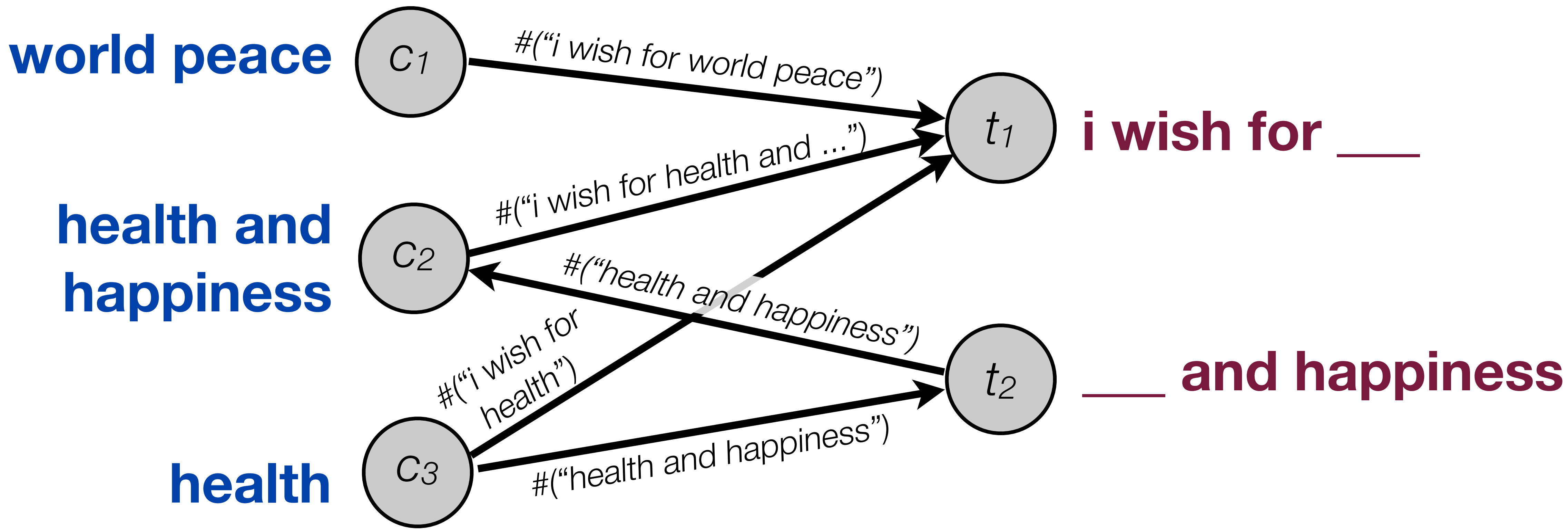
Two kinds of edges:
- $c \rightarrow t$ (weighted by # times content appears in the template)
- $t \rightarrow c$ (weighted by # times template matches a content node)



Ranking template nodes

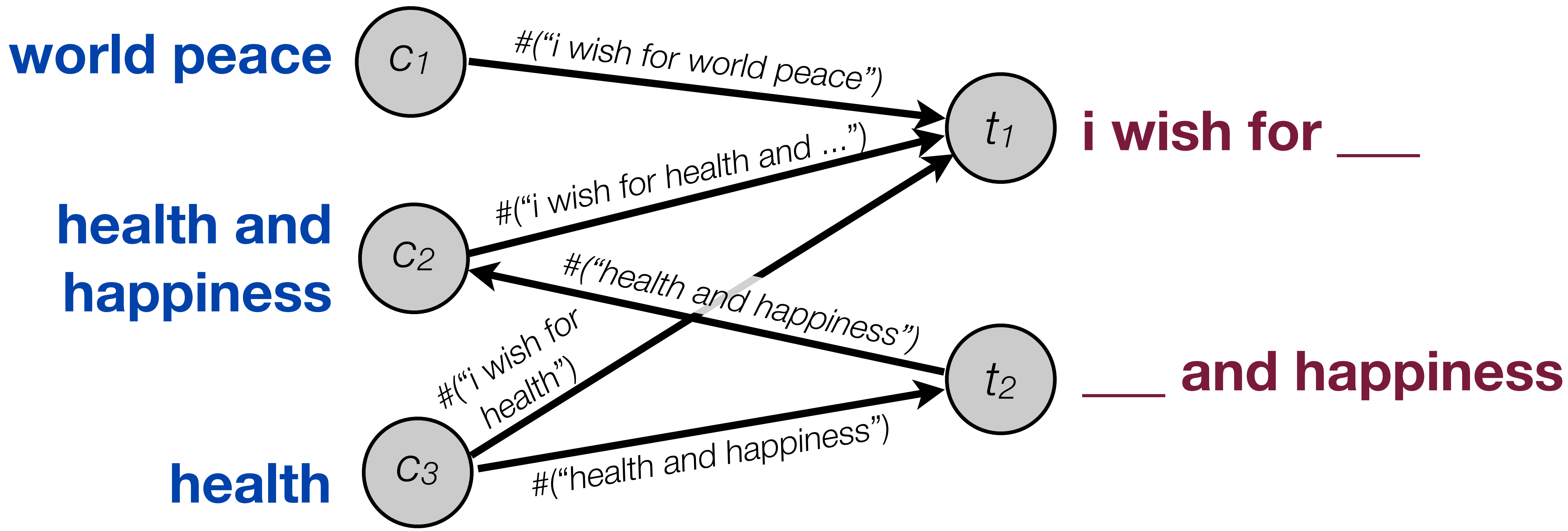


Ranking template nodes



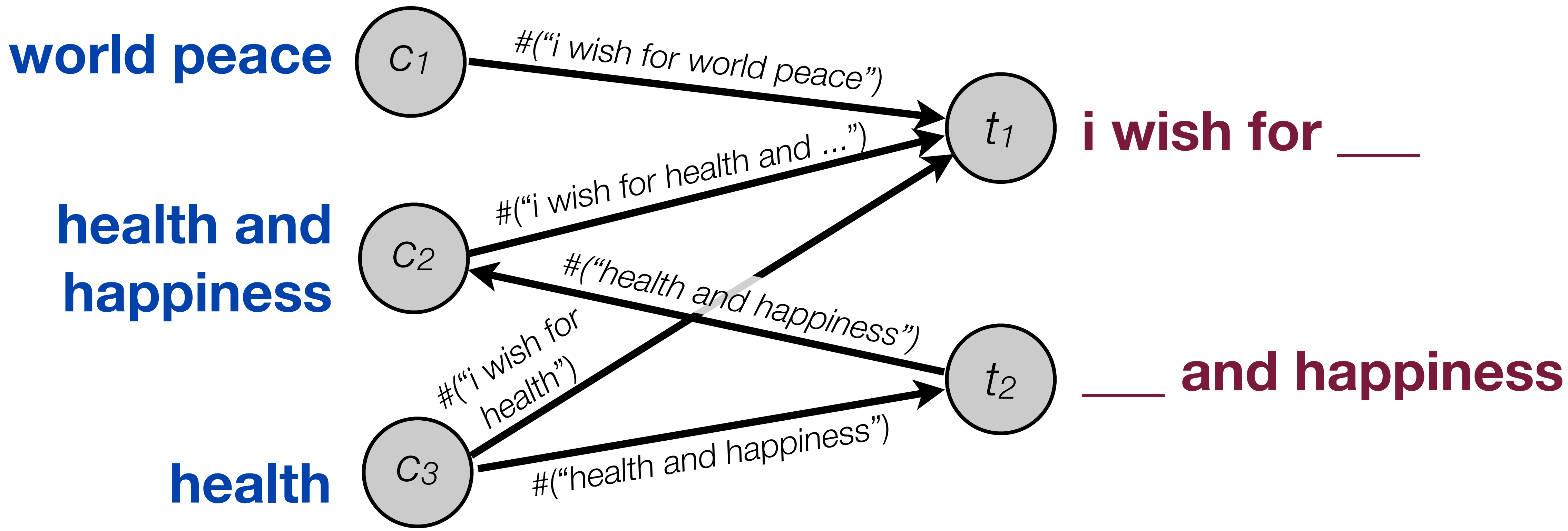
- Useful templates match many complete wishes but few content-only wishes

Ranking template nodes



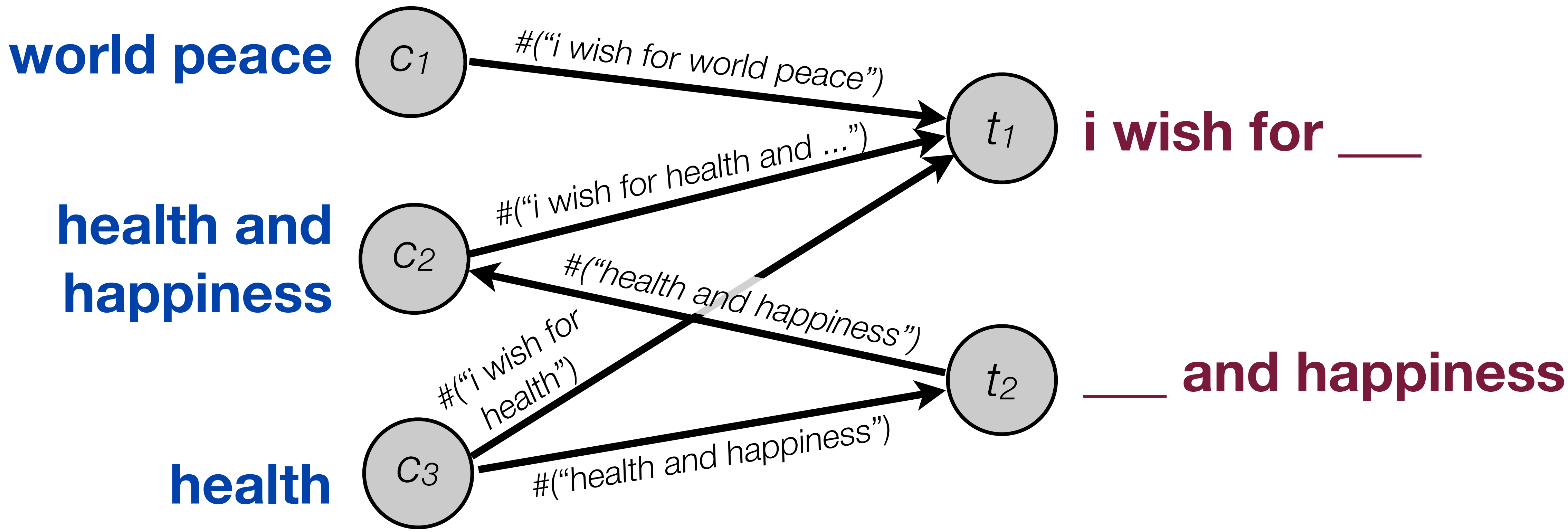
- Useful templates match many complete wishes but few content-only wishes
- We rank all template nodes *t by score(t) = in(t) - out(t)*

Ranking template nodes



- Useful templates match many complete wishes but few content-only wishes
- We rank all template nodes *t by score(t) = in(t) - out(t)*
- Subtracting the out-degree eliminates "bad" templates that contain specific topical content (e.g., "___ and happiness")

Ranking template nodes



- Useful templates match many complete wishes but few content-only wishes
- We rank all template nodes *t by score(t)* = *in(t)* - *out(t)*
- Subtracting the out-degree eliminates "bad" templates that contain specific topical content (e.g., "___ and happiness")
- Apply threshold *score(t)* ≥ *5* to obtain 811 top templates for use as features

Wish template features

Some of the top 811 template features selected by our algorithm

Top 10	Others in Top 200
___ in 2008	i want to ___
i wish for ___	___ for everyone
i wish ___	i hope ___
i want ___	my wish is ___
i want my ___	___ please
___ this year	wishing for ___
i wish ___ in 2008	may you ___
i wish to ___	i wish i had ___
i wish ___ this year	to finally ___
___ in the new year	for my family to have ___

Learning with wish template features

Learning with wish template features

- We use the templates as features for classification in target domains

Learning with wish template features

- We use the templates as features for classification in target domains
- Each template leads to 2 features depending on level of matching in sentence:
  - Whole-sentence match: “**i wish** <u>this mp3 player had more storage</u>”
  - Partial-sentence match: “most of all **i wish** <u>this camera was smaller</u>”

Learning with wish template features

- We use the templates as features for classification in target domains
- Each template leads to 2 features depending on level of matching in sentence:
  - Whole-sentence match: "**i wish** <u>this mp3 player had more storage</u>"
  - Partial-sentence match: "most of all **i wish** <u>this camera was smaller</u>"

- Models using templates:
  - [Templates] uses only these features in a linear SVM
  - [Words+Templates] combines unigram and template features in a linear SVM

Test corpora

Test corpora

- Recall goal of discovering wishes in interesting text domains

Test corpora

- Recall goal of discovering wishes in interesting text domains
- Two test corpora, manually labeled sentences as wish vs. non-wish

Test corpora

- Recall goal of discovering wishes in interesting text domains

- Two test corpora, manually labeled sentences as wish vs. non-wish
  - Consumer product reviews
    - 1,235 sentences from amazon.com and cnet.com reviews (selected from data used in Hu and Liu, 2004; Ding et al., 2008)
    - 12% wishes

Test corpora

- Recall goal of discovering wishes in interesting text domains

- Two test corpora, manually labeled sentences as wish vs. non-wish
    - Consumer product reviews
        - 1,235 sentences from amazon.com and cnet.com reviews (selected from data used in Hu and Liu, 2004; Ding et al., 2008)
        - 12% wishes
    - Political discussion board postings
        - 6,379 sentences selected from politics.com (Mullen and Malouf, 2008).
        - 34% wishes

Test corpora

- Recall goal of discovering wishes in interesting text domains

- Two test corpora, manually labeled sentences as wish vs. non-wish
    - Consumer product reviews
        - 1,235 sentences from amazon.com and cnet.com reviews (selected from data used in Hu and Liu, 2004; Ding et al., 2008)
        - 12% wishes
    - Political discussion board postings
        - 6,379 sentences selected from politics.com (Mullen and Malouf, 2008).
        - 34% wishes

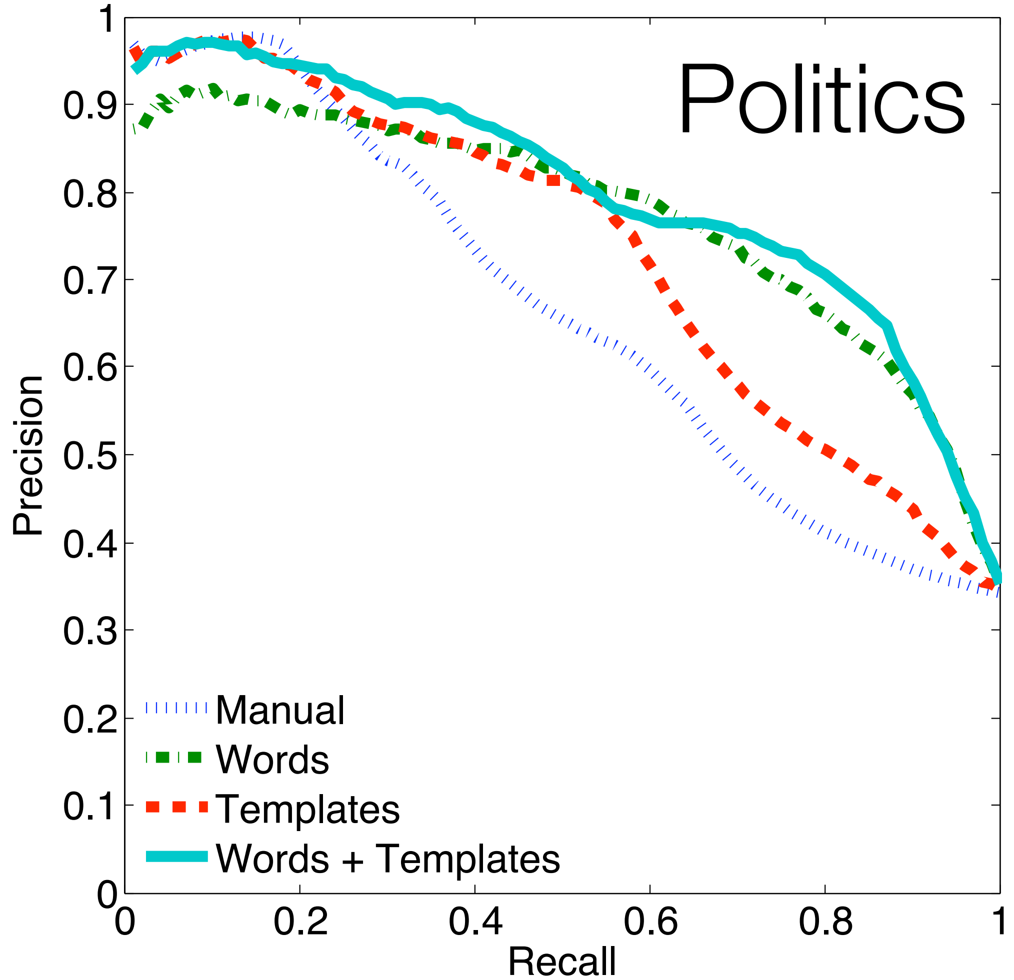
Download from http://pages.cs.wisc.edu/~goldberg/wish_data

Experimental results

10-fold cross validation, linear classifier (SVM^{light} using default parameters)

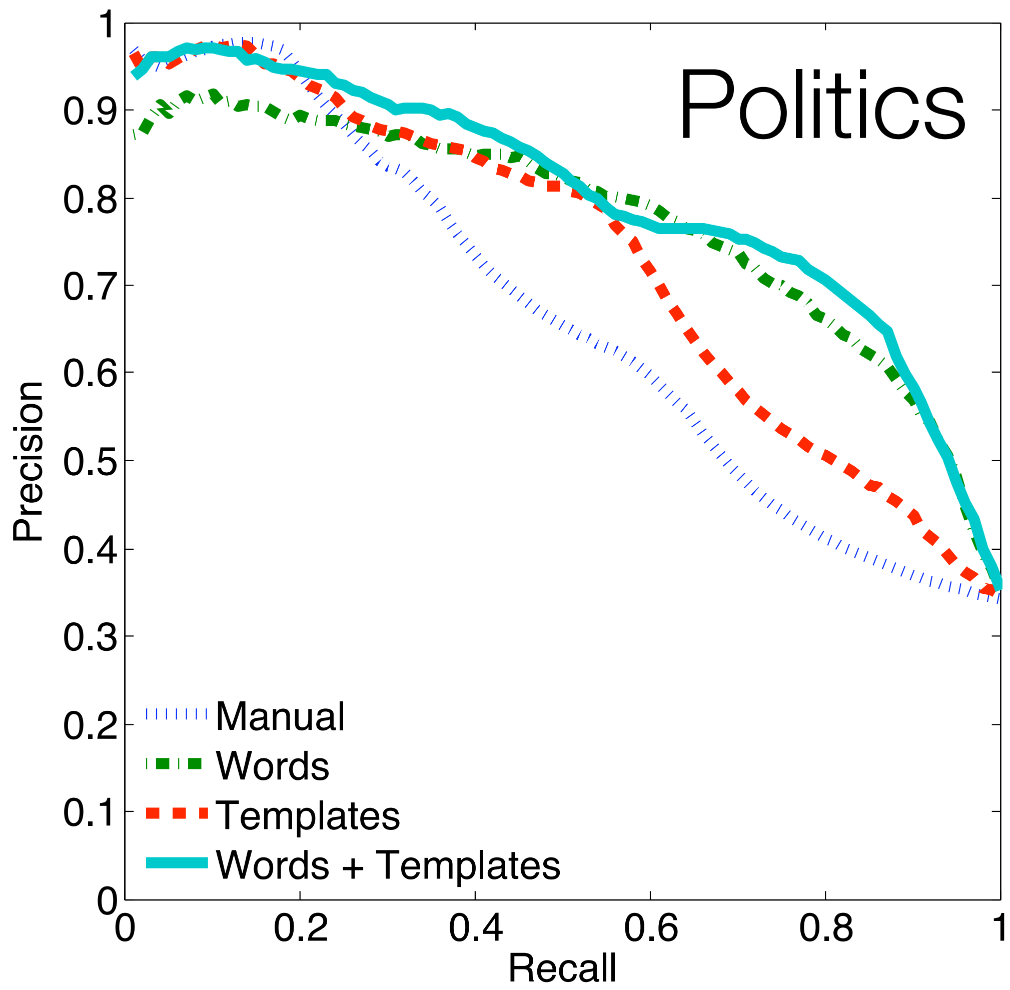
Experimental results

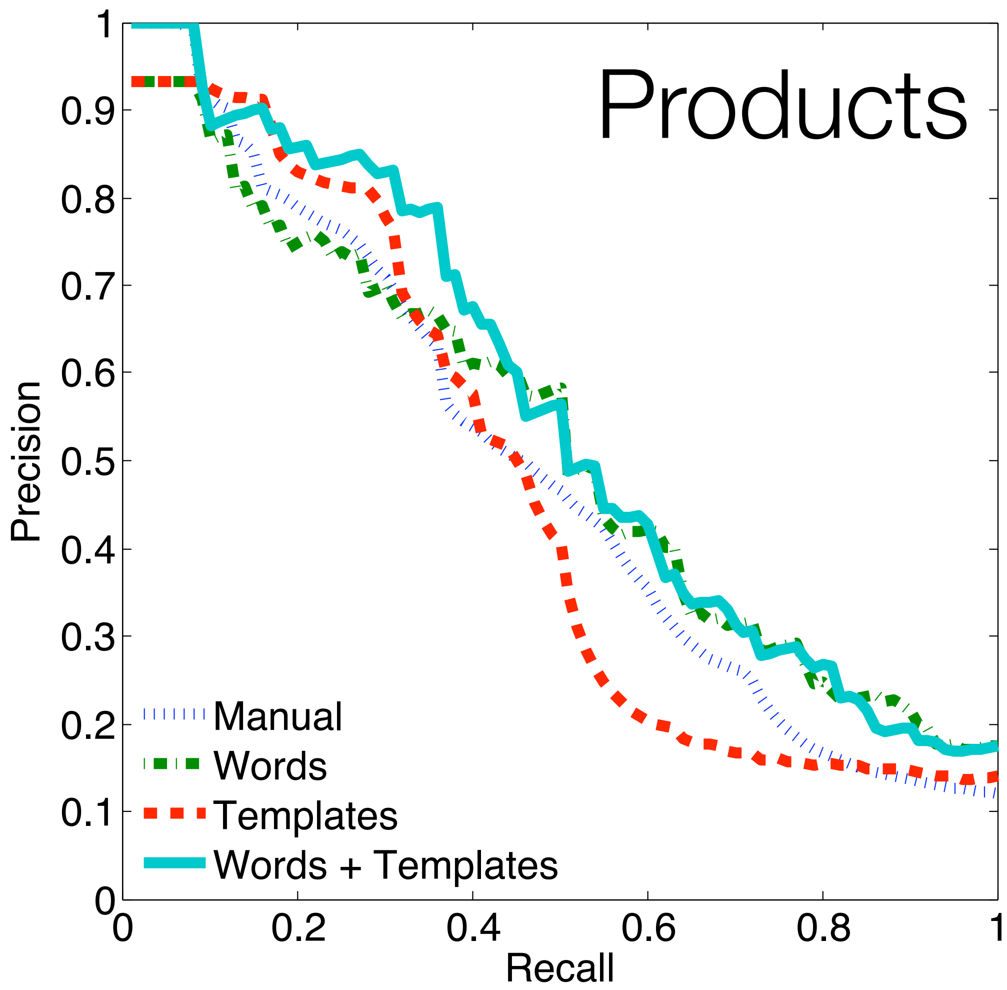
10-fold cross validation, linear classifier (SVM^{light} using default parameters)



Experimental results

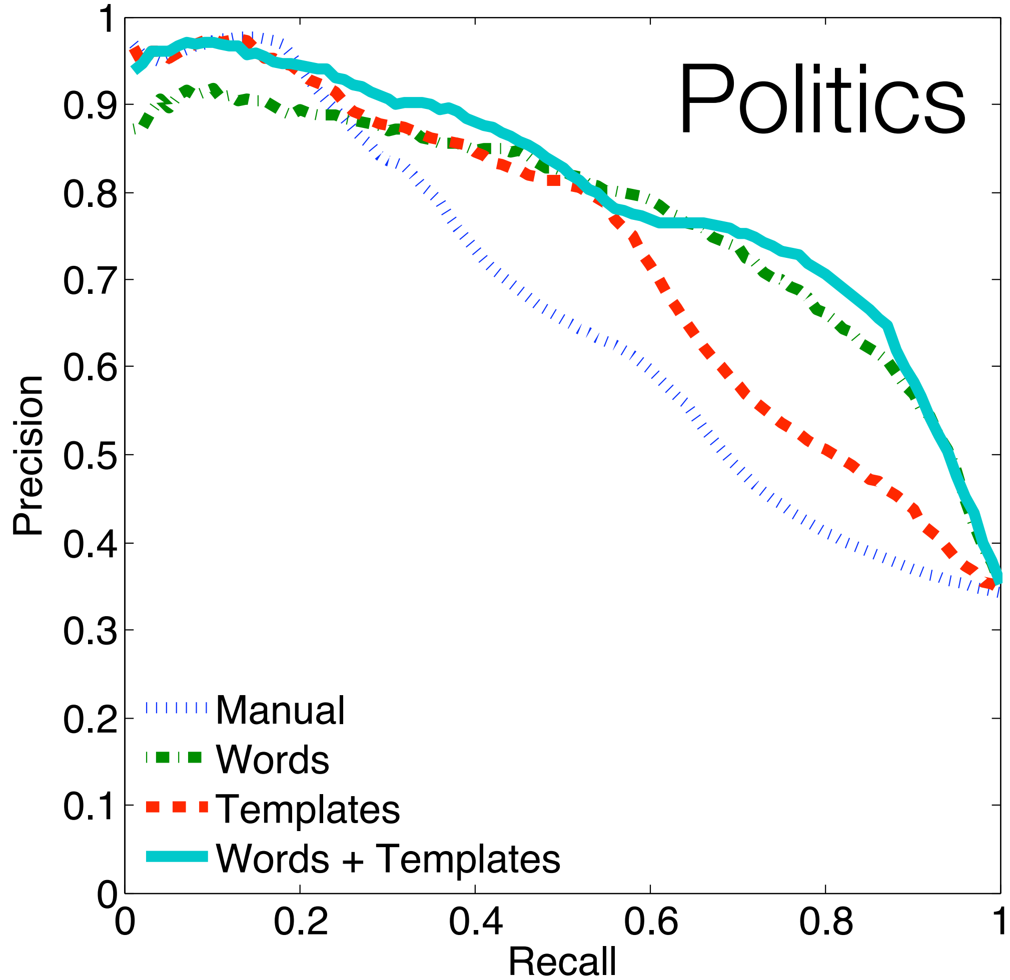
10-fold cross validation, linear classifier (SVM*light* using default parameters)

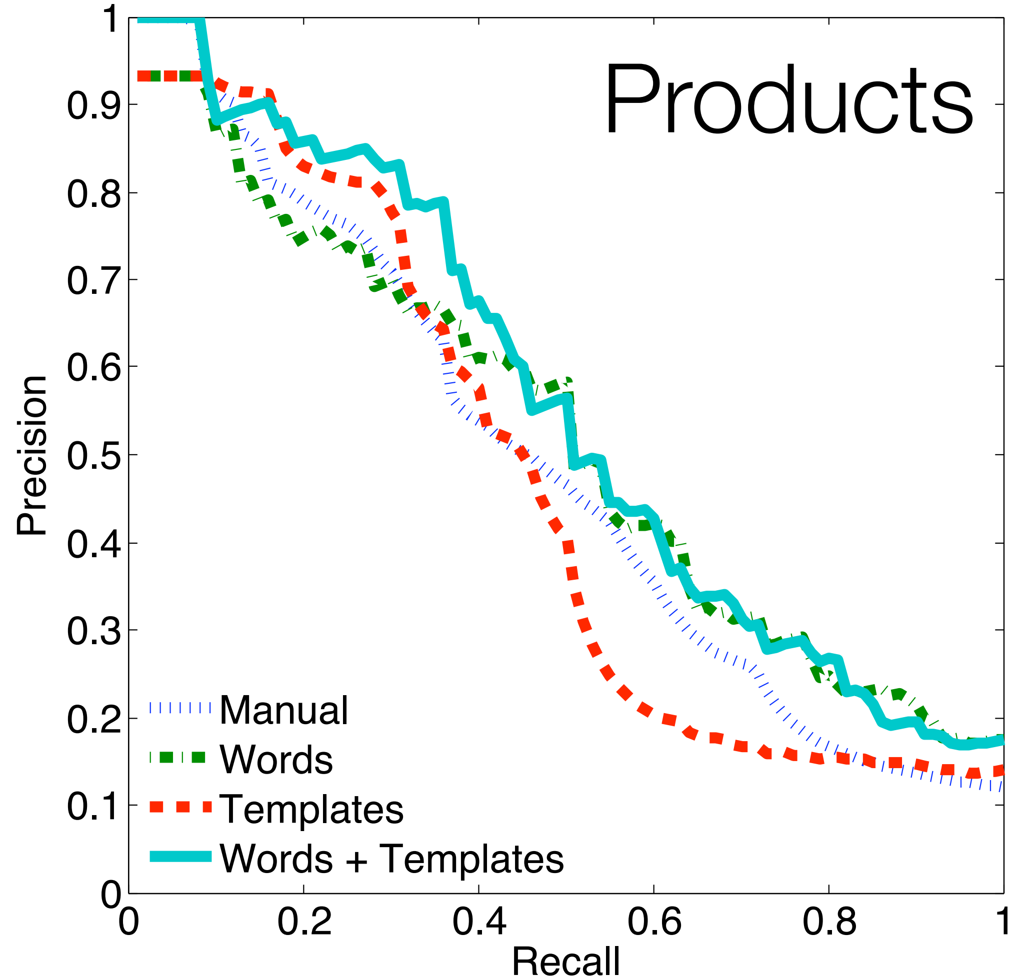




Experimental results

10-fold cross validation, linear classifier (SVM*light* using default parameters)





AUC

Corpus	Manual	Words	Templates	Words + Templates
Politics	0.67 ± 0.03	0.77 ± 0.03	0.73 ± 0.03	0.80 ± 0.03
Products	0.49 ± 0.13	0.52 ± 0.16	0.47 ± 0.16	0.56 ± 0.16

What features are important?

Features with largest magnitude weights for one fold of the Products corpus

Sign	Words	Templates	Words + Templates
+	wish	i hope ___	hoping ___
+	hope	i wish ___	i hope ___
+	hopefully	hoping ___	i just want ___
+	hoping	i just want ___	i wish ___
+	want	i would like ___	i would like ___
-	money	family ___	micro
-	find	___ forever	about
-	digital	let me ___	fix
-	again	___ d	digital
-	you	___ for my dad	you

Conclusions & Future Work

Conclusions & Future Work

- Studied wishes from an NLP perspective for the first time

Conclusions & Future Work

- Studied wishes from an NLP perspective for the first time
- Introduced and analyzed the WISH corpus of ~90,000 wishes

Conclusions & Future Work

- Studied wishes from an NLP perspective for the first time
- Introduced and analyzed the WISH corpus of ~90,000 wishes
- Proposed new wish detection task and simple baselines

Conclusions & Future Work

- Studied wishes from an NLP perspective for the first time
- Introduced and analyzed the WISH corpus of ~90,000 wishes
- Proposed new wish detection task and simple baselines
- Generated wish templates that transfer to new domains

Conclusions & Future Work

- Studied wishes from an NLP perspective for the first time
- Introduced and analyzed the WISH corpus of ~90,000 wishes
- Proposed new wish detection task and simple baselines
- Generated wish templates that transfer to new domains
- Built wish-annotated test corpora in product review and political domains

Conclusions & Future Work

- Studied wishes from an NLP perspective for the first time
- Introduced and analyzed the WISH corpus of ~90,000 wishes
- Proposed new wish detection task and simple baselines
- Generated wish templates that transfer to new domains
- Built wish-annotated test corpora in product review and political domains

- Much future work in wish detection remains:
  - Additional wish-sensitive features
  - Annotated training data is expensive → semi-supervised learning

Acknowledgements

We'd like to thank:

Times Square Alliance for providing the WISH corpus

Wisconsin Alumni Research Foundation

Yahoo! Key Technical Challenges Program

&

you!

Download test corpora at
http://pages.cs.wisc.edu/~goldberg/wish_data