Text Normalization Using Lexical and Contextual Features

M.S. Thesis
by
Çağıl Uluşahin Sönmez
Advisor : Arzucan Özgür

Jan 14, 2014

Motivation

Its a btf nite, lukin for smth fun to do,
I think I wanna be w ma frnds.

Its a beautiful night, looking for something fun to do,
I think I want to be with my friends.

Text Normalization cont.

- In Vocabulary(IV) words:
- **Out Of Vocabulary(OOV) words**
- Ill-formed words

Imgine a world **wer googling smt** about html or related stuff does not send **u** to **w3schools**.	Imagine a world where **googling** something about html or related stuff does not send you to **w3schools**.
rozis r red **vilitz r blu** **sunflowurs r yelo** **u wer probly ekspektin sumthin romentik** but **deez r jus gardenin fakts**	rozes are red violets are blue sunflowers are yellow you were probably expecting something romantic but these are just gardening

Text Normalization

Two steps: • Detection • Normalization

Its a btf nite, lukin for smth fun to do, I think I **wanna** be w ma frnds.	It's a beautiful night, looking for something fun to do, I think I **wanna** be with my friends.
Dnt always follow da crowd,stand 4 wat u blv in.	Don't always follow the crowd, stand for what you believe in.
Work f a cos, not for applause. Live life to exprss, not to imprss :)	Work for a cause, not for applause. Live life to express, not to impress :)
There r sm songs u dont want 2 listen 2 yl walking cos when u start dancing ppl won't knw y.	There are some songs you don't want to listen to while walking because when you start dancing people won't know why.

Related Work

- Brill and Moore, 2000 proposed a novel noisy channel model for spell checking based on string to string edits
- Toutanova et al., 2002 extended this error model with phonetic similarities over words
- Aw et al., 2006 proposed a phrase-based statistical machine translation (MT) model
- Choudhury et al., 2007 proposed a supervised Hidden Markov Model based approach

Related Work cont.

- More recent approaches handled the text normalization task by building normalization lexicons
- Han et al., 2011 presents a normalization lexicon using lexical features and contextual features of OOV words
- Gouws et al., 2011 - highly dependent on user centric information such as the geological location of the users and the twitter client that the tweet is received from
- Pennel and Liu, 2011 used a character level MT system
- Liu et al., 2012 integrates human perspective modelling (an extended noisy channel model)

Related Work cont.

- Yang and Eisenstein, 2013 introduced an unsupervised log linear model for text normalization
- Their joint statistical approach uses local context based on language modeling and surface similarity
- Hassan and Menezes, 2013 generated a normalization lexicon using Markov random walks on a contextual similarity lattice

A Graph Based Approach for Contextual Text Normalization

- A Text normalisation system **based on** Word Association Graph
- **Unsupervised**, no need for labeled data
- Uses **input context** & corpus based contextual information

Imgine a world **wer** googling smt	Imagine a world **where** googling something
u **wer** probly ekspektin sumthin	you **were** probably expecting something

- Best precision and f-measure with good recall

Our Methodology

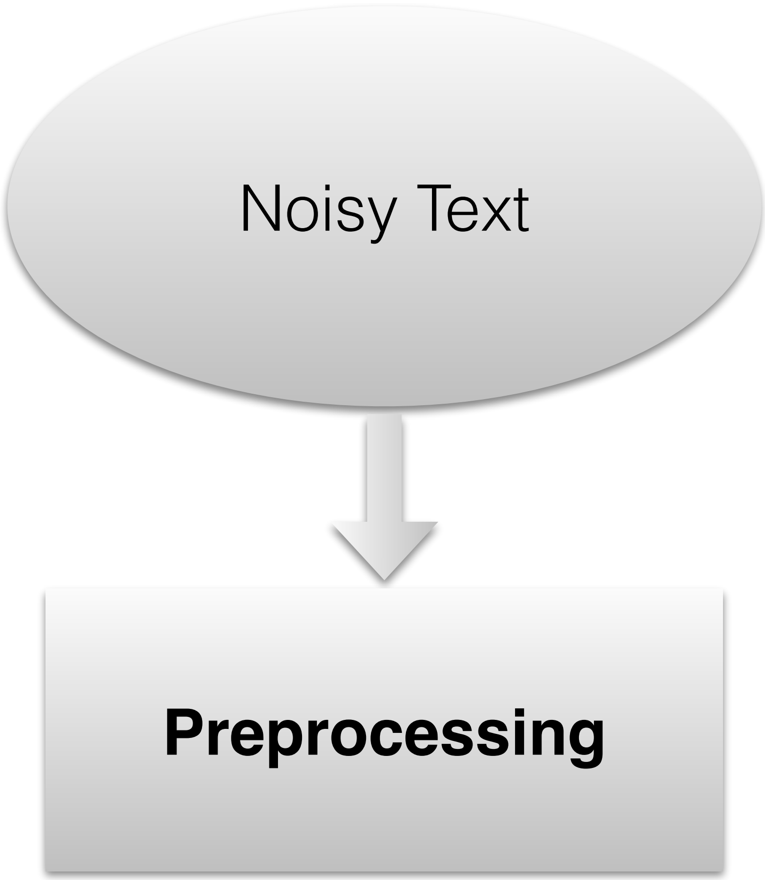
- graph based approach
- social text (twitter)
- contextual and lexical similarity features
- A slang dictionary used as an external resource

Preprocessing

- Preprocessing:

  i) Tokenization

  ii) Part of Speech(POS) tagging



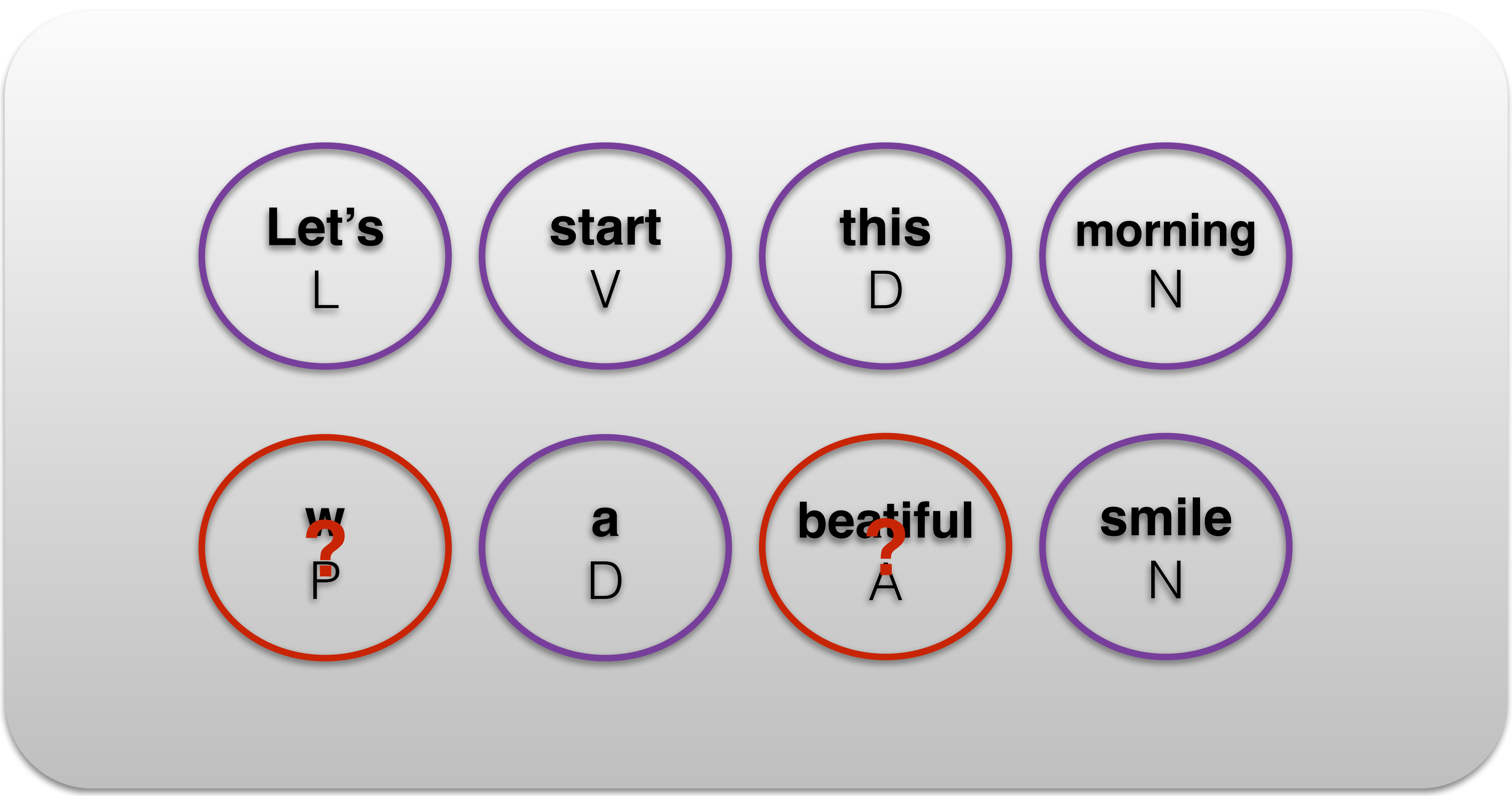
Preprocessing cont.

Dnt always follow da crowd,stand 4 wat u blv in.

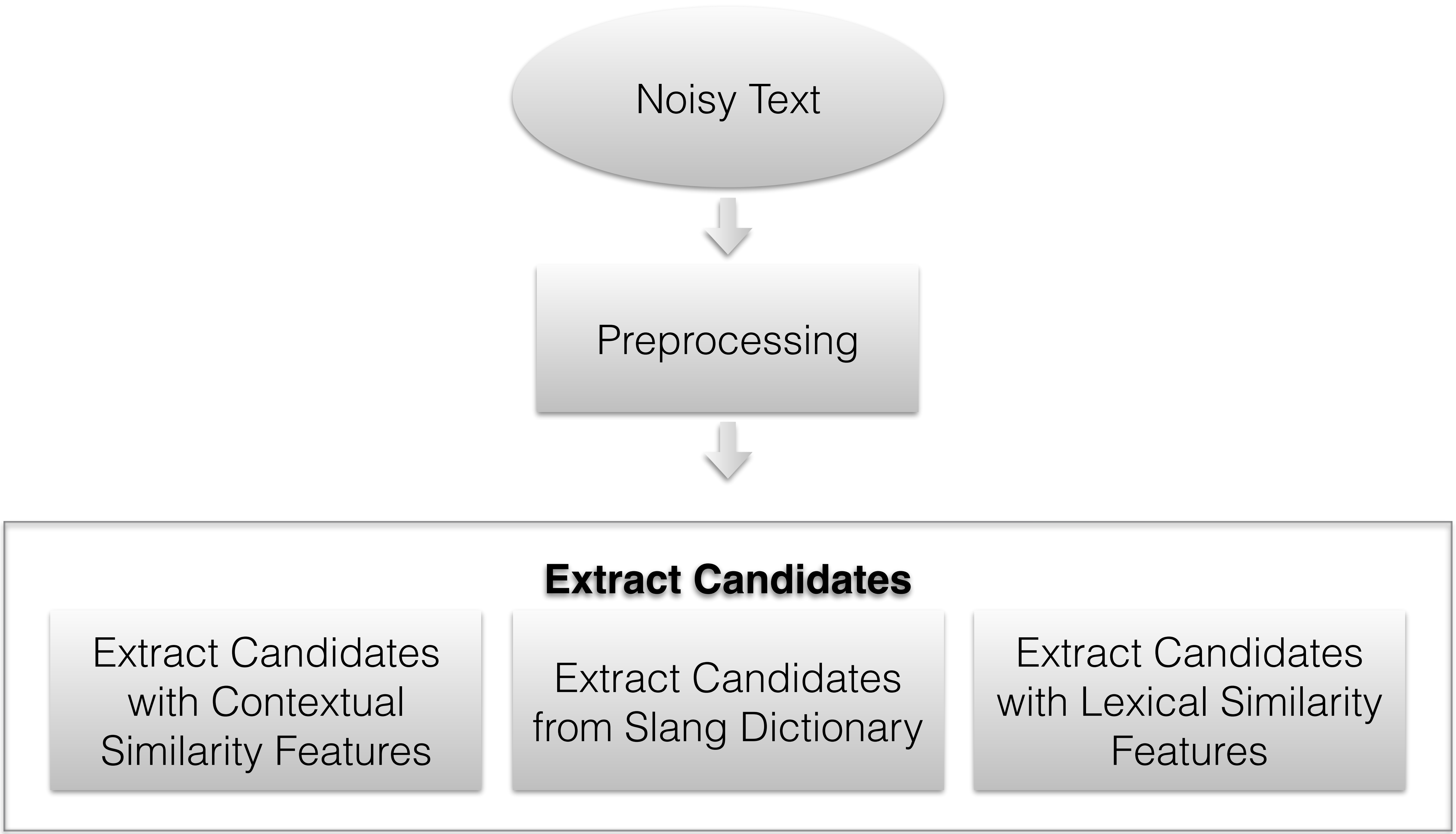
Dnt	Verb	.91
always	Adverb	.98
follow	Verb	.99
da	Determiner	.98
crowd	Noun	.99
,	Punctuation	.99
stand	Verb	.84
4	Preposition	.61
wat	Pronoun	.93
u	Pronoun	.99
blv	Verb	.97
in	Preposition	.92
.	Punctuation	.98

Extracting Candidates

- find normalisation candidates for each OOV word in the input text

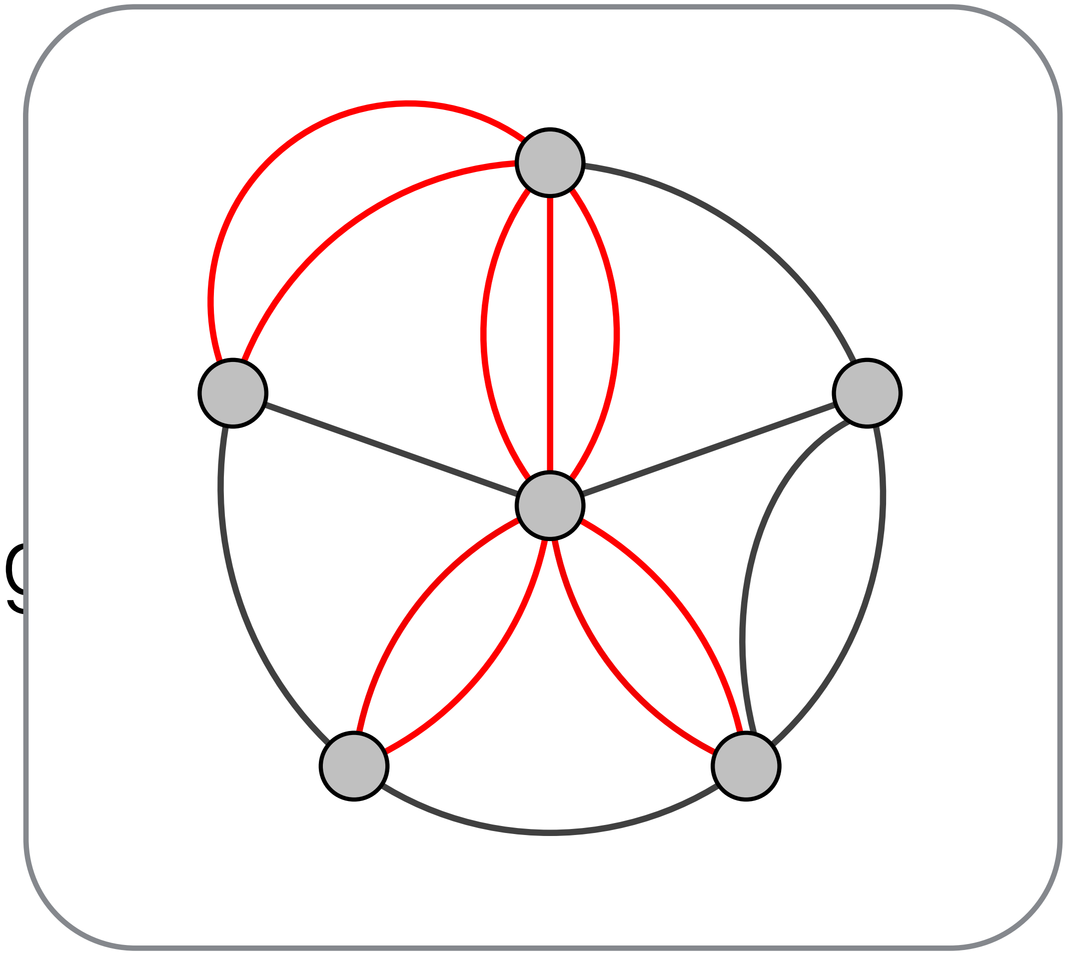


Extracting Candidates

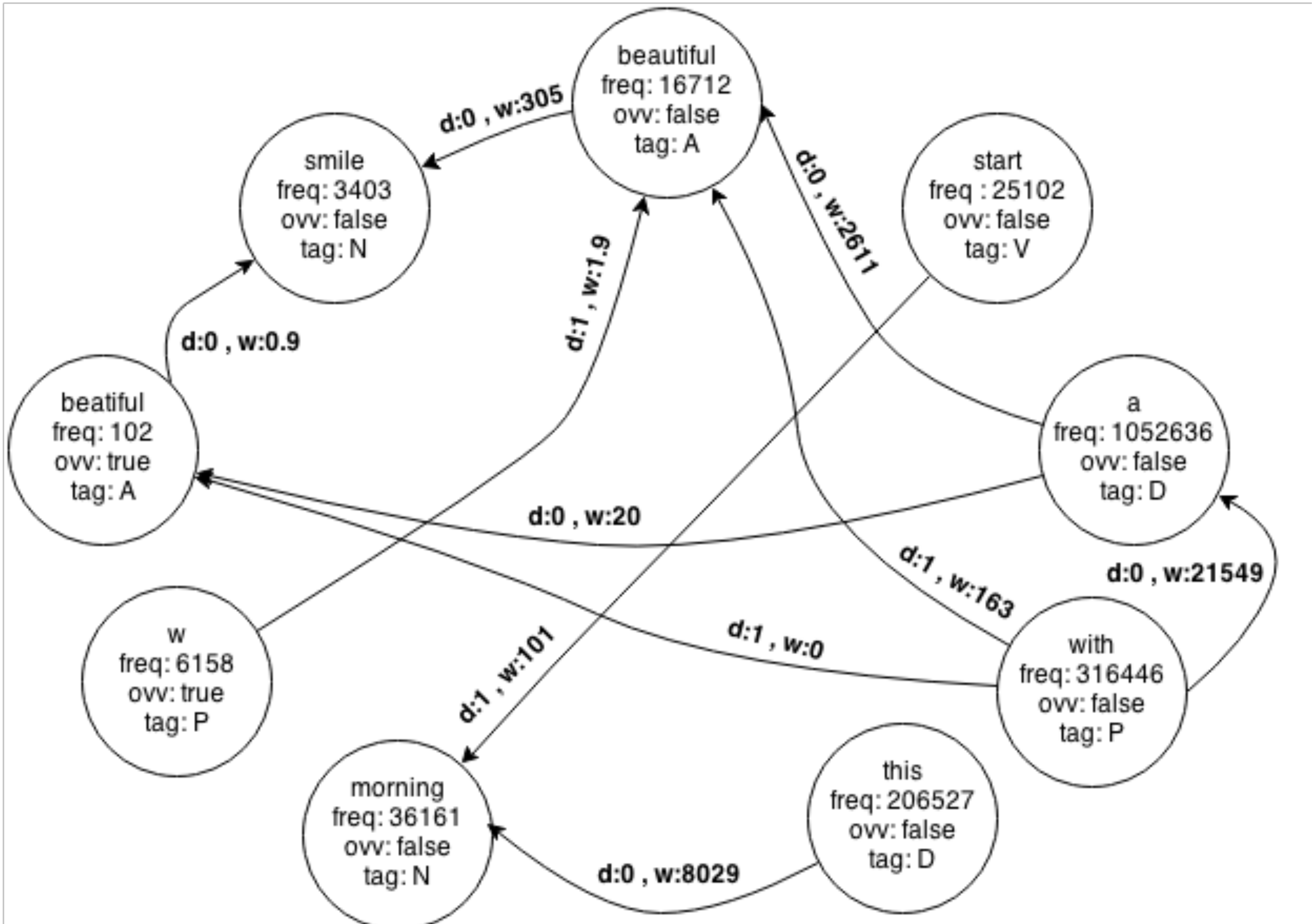


Extracting Candidates with Contextual Similarity Features

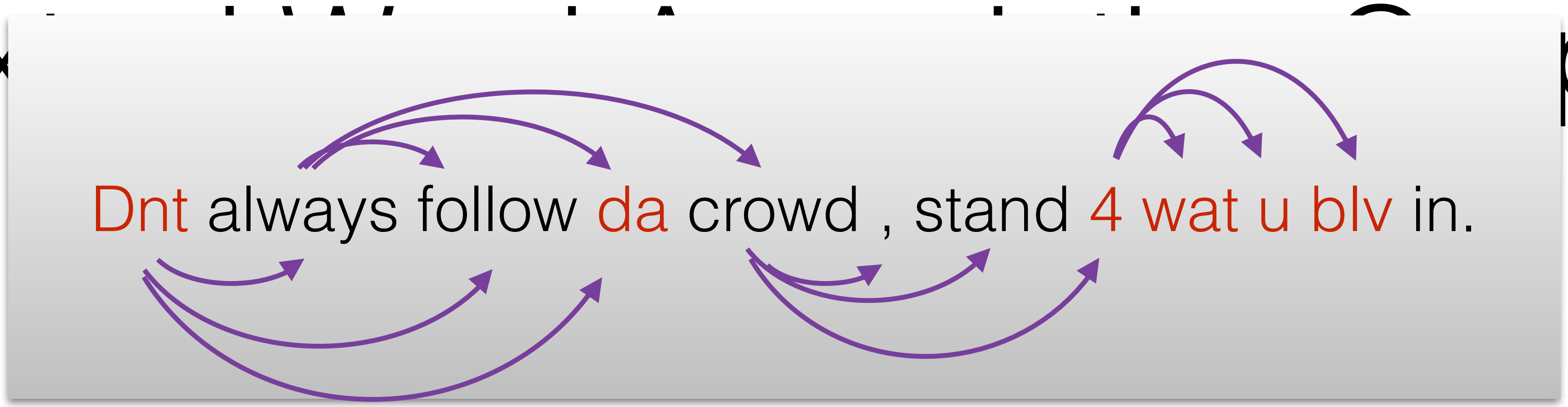
- Contextual Word Association Graph (CWA-graph)
- directed weighted multigraph
- models contextual information
- relative positions of the POS tag



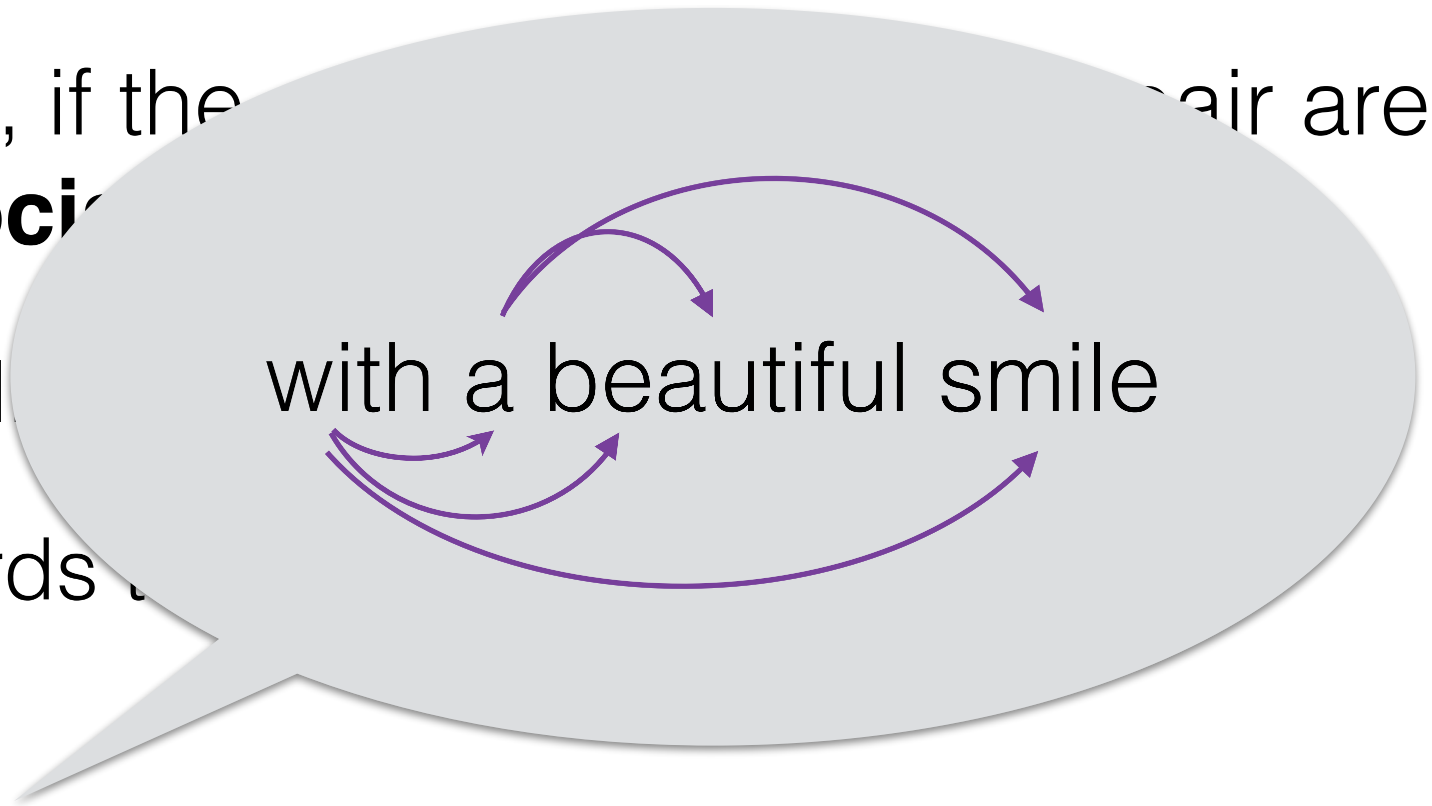
Contextual Word Association Graph



Contex ph

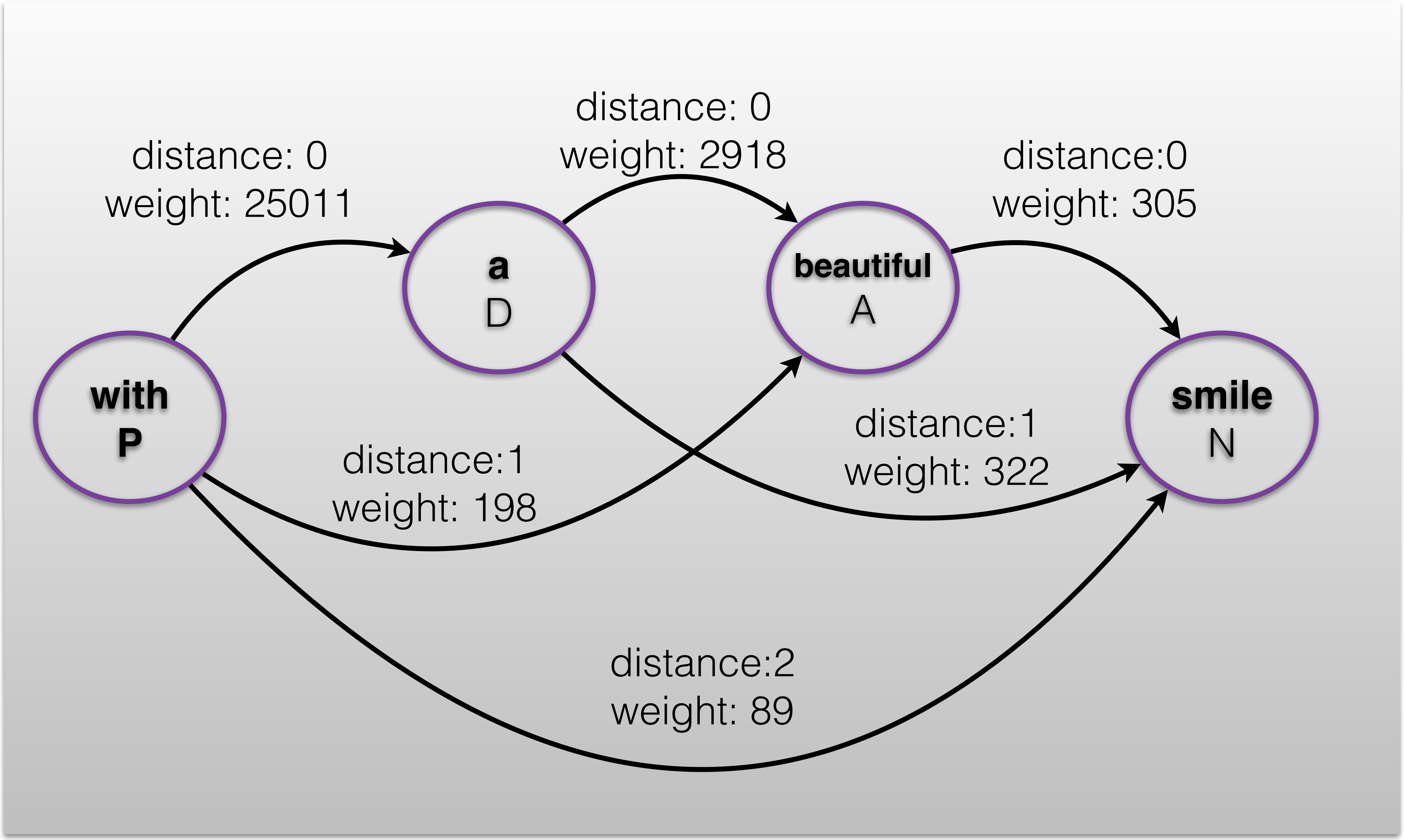


- Edges are created, if the air are **contextually associ**
  - requires a word d
  - requires both words threshold

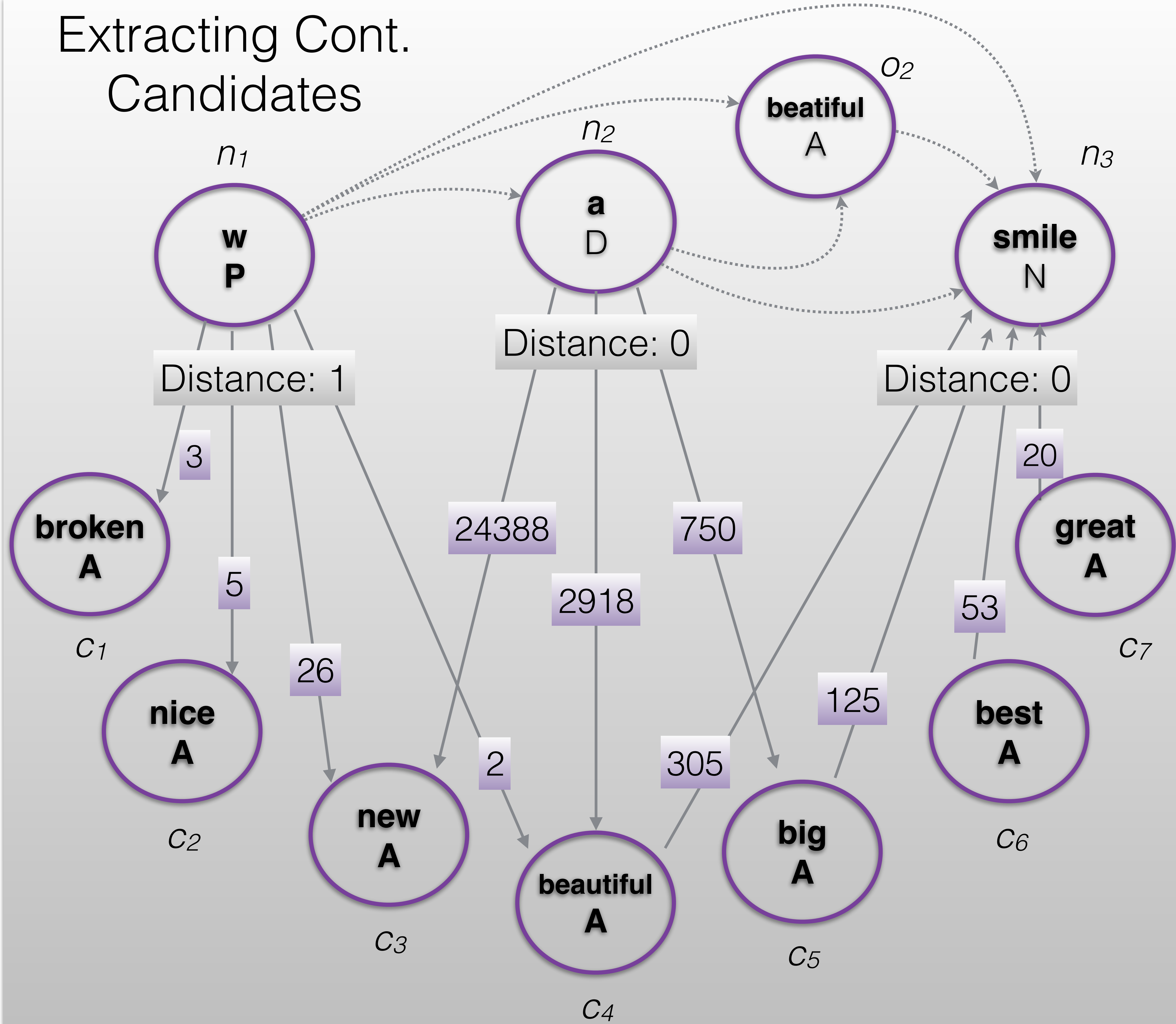


- directionality: based on the sequence of words
- The direction and the distance: a unique triplet

Distance and Edge Weight



Extracting Cont. Candidates



Extracting Candidates with Lexical Similarity Features

1. find new candidates (lexically similar)

2. filter the candidates (edit-d and phonetic-d thresholds)

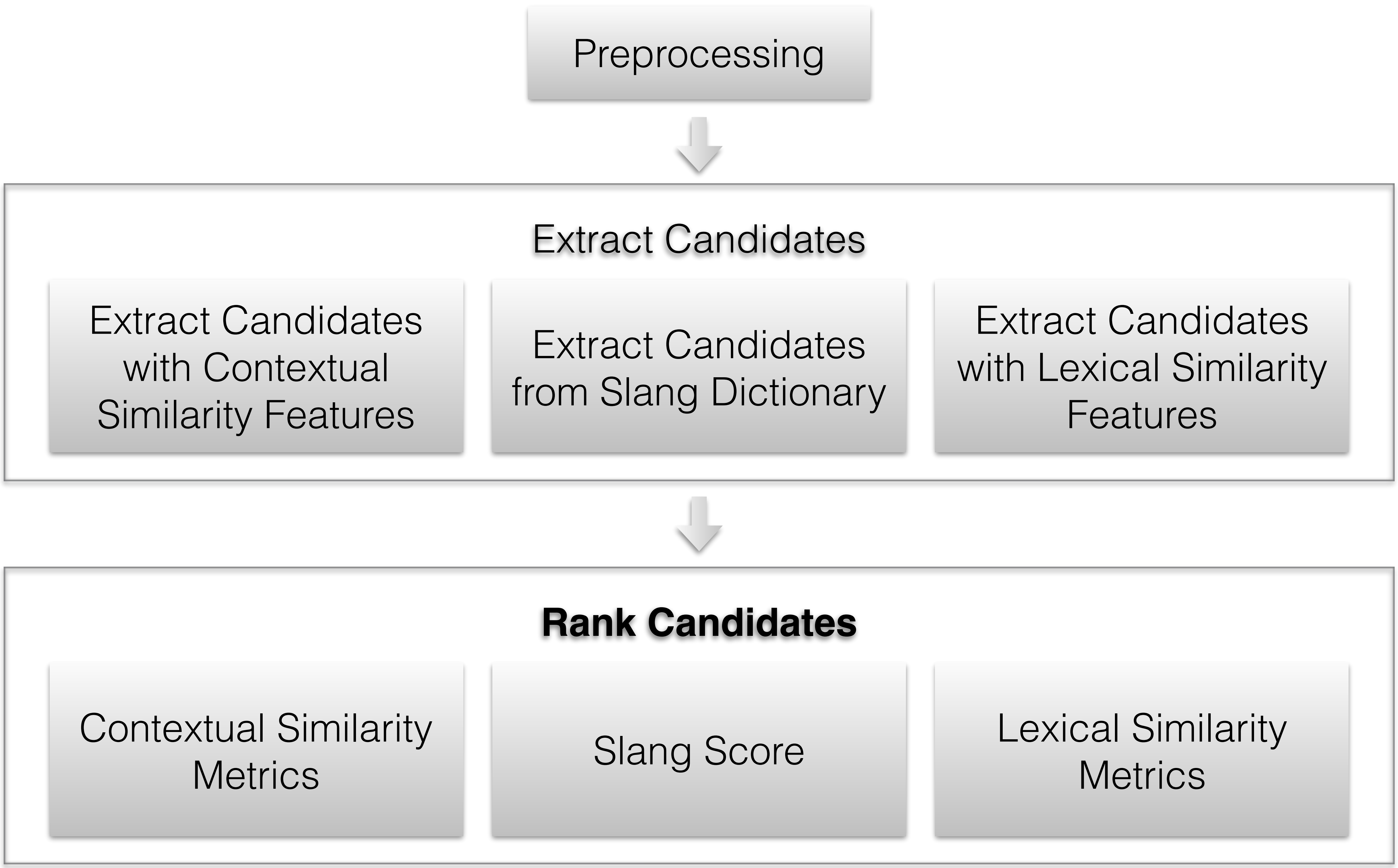
    - edit distance

    - double metaphone (phonetic edit distance)

Lexically Similar Candidates

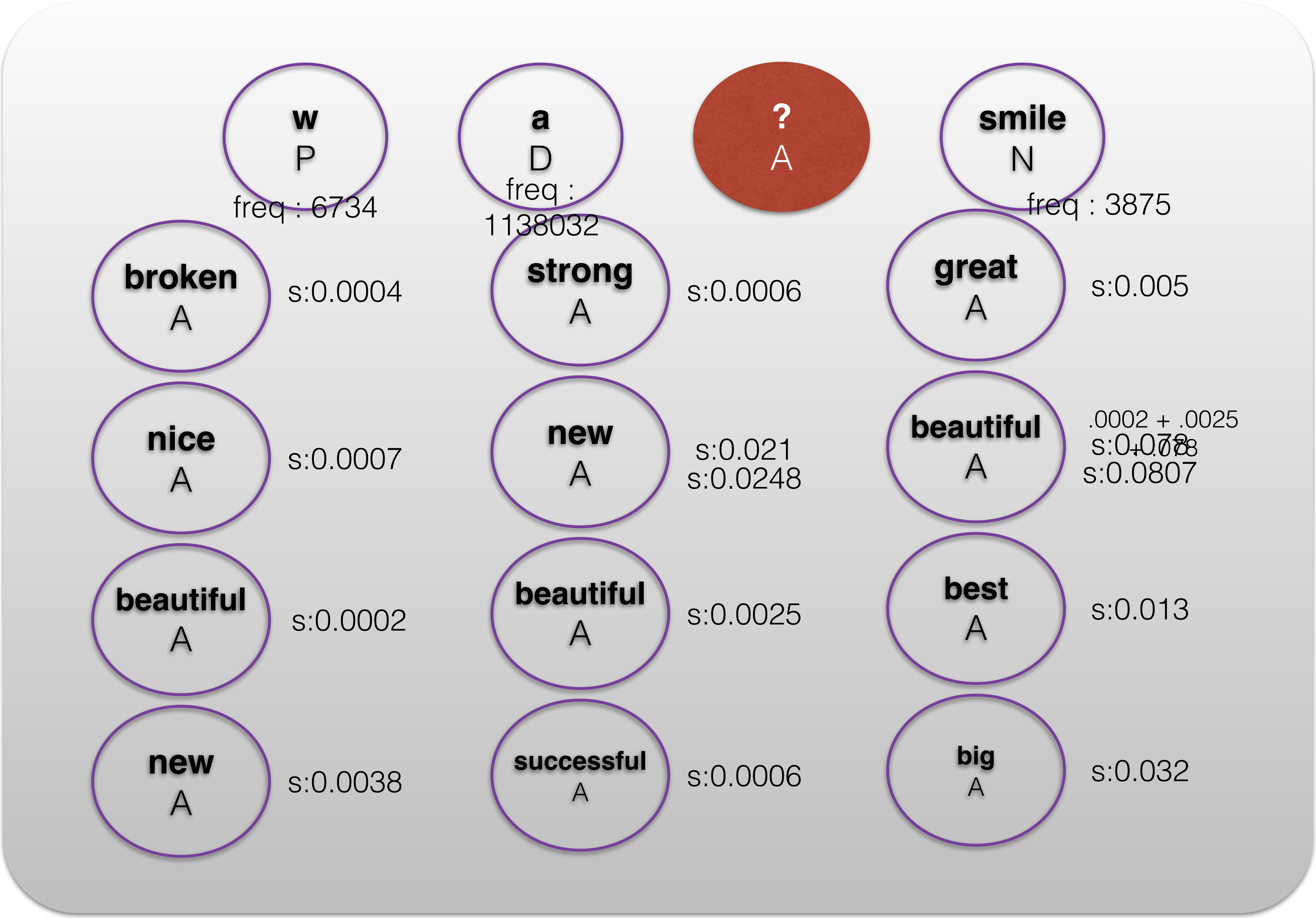
OOV	Candidate	Edit Distance	Phonetic Distance
missin (MSN)	missing (MSNK)	1	1
missin (MSN)	missed (MST)	2	1
confrims (KNFR)	confirms (KNFR)	2	0
confrims (KNFR)	confirm (KNFR)	**3**	0
soemthing (SM0N,SMTN)	something (SM0N,SMTN)	2	0
soemthing (SMTN)	sorting (SRTN)	**3**	1
smt (SMT,XMT)	something (SMTN)	**6**	1

Ranking Candidates



Contextual Similarity Metrics

- Edge Weight Score
    1. related to many neighbours
    2. have a high association score with each neighbour
- Frequency Score
    - a real number between 0 and 1
    - proportional to the frequency of the candidate within the corpus


w
P
freq : 6734
a
D
freq :
1138032
?
A
smile
N
freq : 3875
broken
A
s:0.0004
strong
A
s:0.0006
great
A
s:0.005
nice
A
s:0.0007
new
A
s:0.021
s:0.0248
beautiful
A
.0002 + .0025
s:0.078
+.078
s:0.0807
beautiful
A
s:0.0002
beautiful
A
s:0.0025
best
A
s:0.013
new
A
s:0.0038
successful
A
s:0.0006
big
A
s:0.032

Candidates with Edge Weight Score and Frequency Score

beautiful
freq: 17900
A

edgeWeightScore = 0.18
frequencyScore = 1
contextSimScore = (1 * 0.18) + (0.5 * 1)
= **0.679**

big
freq: 191713
A

edgeWeightScore = 0.12
frequencyScore = 1
contextSimScore = **0.62**

new
freq: 36252
A

edgeWeightScore= 0.02
frequencyScore = 1
contextSimScore = **0.52**

Lexical Similarity Metrics & Slang Score

OOV	Candidate	LCSR Score	Edit-Dist Score	Slang Score
missin (MSN)	missing (MSNK)	0.8571	0.8572	0
missin (MSN)	missed (MST)	0.6667	0.6666	0
confrims (KNFR)	confirms (KNFR)	0.8750	0.75	0
confrims (KNFR)	confirm (KNFR)	0.7500	0.6240	0
soemthing (SMTN)	something (SMTN)	0.8889	0.7778	0
soemthing (SMTN)	sorting (SRTN)	0.6666	0.6666	0
smt (SMT)	something (SMTN)	0.3333	0.3333	1

Final Ranking of Candidates

OOV	Candidate	edgeWeight Score	freq Score	LCSR Score	Edit-Dist Score	Slang Score	Final Score
follwers	followers	0.0505	1	0.8888	0.8888	0	1.8839
follwers	follower	0.0481	1	0.8750	0.7500	0	1.7981
follwers	flowers	0.0182	1	0.7500	0.7500	0	1.6432
follwers	follower's	0.1799	0.2	0.8000	0.8000	0	1.4799
follwers	flower	0.0248	1	0.6250	0.6250	0	1.4623
follwers	dollars	0.0084	1	0.6250	0.6250	0	1.4459

Experiments

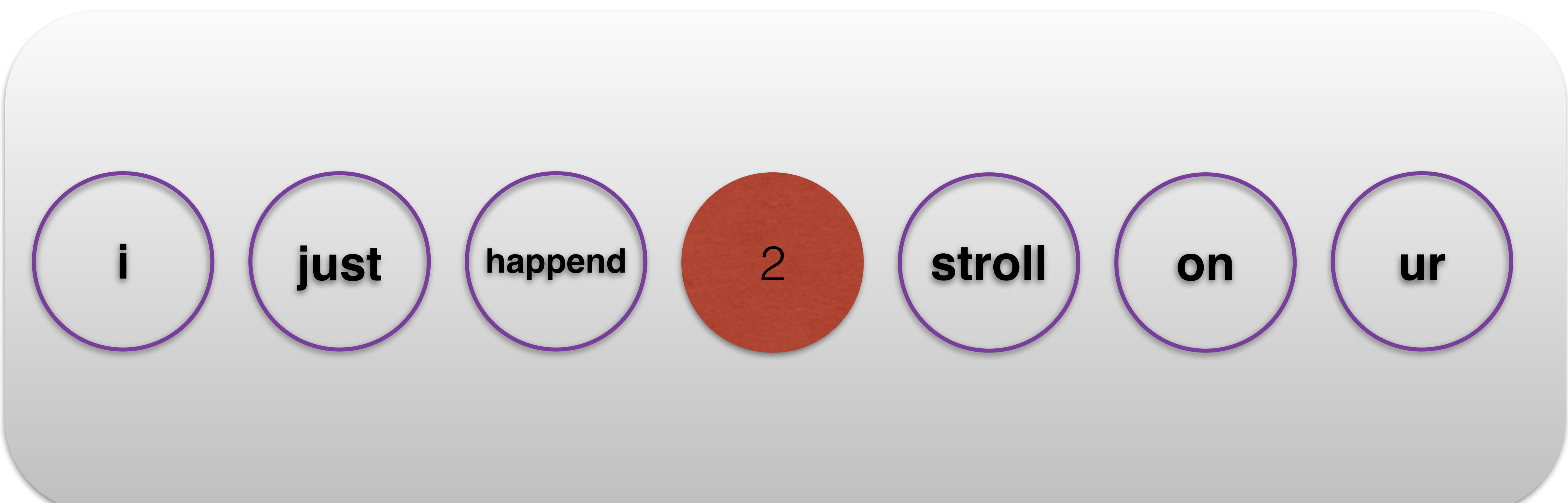
- Graph Generation: We extracted 1 GB of English tweets from **Stanford's 476 million Twitter Dataset**
- POS tagger: **CMU Ark Tagger**, which is a social media specific POS tagger achieving an accuracy of 95% over social media text.
- We only kept the nodes with a minimum frequency of 9.
- The resulting graph contains 105428 nodes and 46609603 edges.
- While extending the candidate set with lexical features we use $threshold_{edit} \leq 2 \vee threshold_{phonetic} \leq 1$ to keep up with the settings in Han et al.

Experiments cont.

- First Dataset: **LexNorm1.1**
  - 549 tweets with 1184 manually annotated ill-formed OOV tokens
- Second Dataset: **Pennell Trigram dataset**
  - 985 trigrams from1925 sentences and 985 manually annotated ill-formed OOV tokens
  - **SMS-like Corpus**: collected using only messages sent via SMS

Window Size

- The window size is chosen as 7, with 3 neighbours in each side of the OOV token. (when available)
- Ex: "*I just <u>happend</u> 2 stroll on <u>ur</u> name saw a twit <u>pic</u> I liked so <u>w</u> not <u>u</u> know keep it beautiful :) ? ? thank <u>u</u> !*"



Results Using Different Window Sizes

Window Size	Precision	Recall	F-measure
3	85.30	79.00	82.00
5	85.60	79.10	82.20
7	**85.50**	**79.20**	**82.20**
9	85.20	79.00	82.00
n	85.20	79.00	82.00

System Tuning

Threshold	Precision	Recall	F-measure
≤ 1	81.2	80.8	81
1.1	81.5	80.8	81.2
1.2	82.2	80.7	81.4
1.3	83.7	80.2	81.9
1.4	84.2	80.0	82.0
1.5	85.5	79.2	82.2
1.6	88.8	75.1	81.4
1.7	91.1	72.8	80.9
1.8	92.3	67.6	78
2	94.1	56.4	70.5

Results on LexNorm1.1

Method	Precision	Recall	F-measure
Han & Baldwin,2011	75.30	75.30	75.30
Liu et al., 2011	84.13	78.38	81.15
Hassan et al., 2013	85.37	56.40	69.93
Yang et al., 2013	82.09	**82.09**	82.09
CWA-Graph	**85.50**	79.20	**82.20**

Results on Trigram Dataset

Method	Precision	Recall	F-measure
Pennell and Liu,2011	69.70	**69.70**	69.70
CWA-Graph	**78.20**	68.5	**73.10**

Future Work

- OOV Detection
- Turkish Text Normalization
- Analysing different graph sizes

Summary of Contributions

- an **unsupervised** text normalization approach
- utilizes **lexical**, **contextual** and **grammatical** features of social text
- a **novel** graph based system
- state of the art precision and f-score
- can be tuned to achieve very high precisions without sacrificing much from recall