NANYANG TECHNOLOGICAL UNIVERSITY

Product Name Recognition and Normalization in Internet Forums

Yangjie Yao **Aixin Sun**

Nanyang Technological University

Singapore

Users’ feedback on products in Internet forums

Name variation	#users	Name variation	#users
1. galaxy s3	553	14. lte s3	46
2. s3 lte	343	15. galaxy s3 lte	45
3. samsung galaxy s3	284	16. s3 non lte	32
4. s iii	242	17. samsung galaxy siii	32
5. galaxy s iii	225	18. sgs 3	27
6. samsung s3	219	19. samsung galaxy s3 lte	22
7. sgs3	187	20. sg3	21
8. siii	149	21. gsiii	16
9. samsung galaxy s iii	145	22. samsung galaxy s3 i9300	15
10. i9300	120	23. samsung i9300 galaxy s iii	13
11. gs3	82	24. s3 4g	11
12. galaxy siii	61	25. 3g s3	11
13. i9305	52	–	

Samsung Galaxy SIII (LTE and Non-LTE versions)

Our target by examples

1. True, **Desire** [HTC Desire] might be better if compared to **X10** [Sony Ericsson Xperia X10] but since I am using **HD2** [HTC HD2], it will be a little boring to use back HTC ...
2. I just wanna know what problems do users face on the **OneX** [HTC One X]... of course I know that knowing the problems on **one x** [HTC One X] doesn't mean knowing the problems on **s3** [Samsung Galaxy SIII]
3. Still prefer **ip 5** [Apple iPhone 5] then **note 2** [Samsung Galaxy Note II]...
4. oh, the mono rich recording at **920** [Nokia Lumia 920] no better than stereo rich recording at **808** [Nokia 808 PureView].

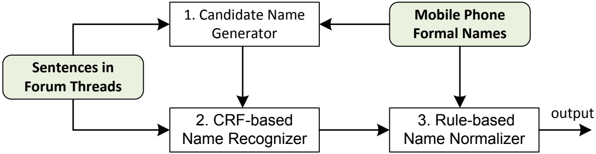
Our approach: generate, recognize, normalize

Input

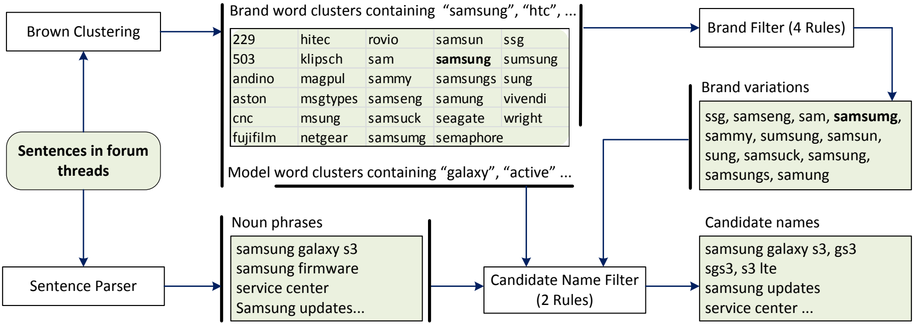
- Posts or messages from domain-relevant Internet forums
- List of formal names

Approach

- Generate candidate names based on naming convention
- Recognize true product names from candidate names
- Normalize names based on naming convention


1. Candidate Name Generator
Mobile Phone Formal Names
Sentences in Forum Threads
2. CRF-based Name Recognizer
3. Rule-based Name Normalizer
output

Candidate name generation



Brown Clustering
Sentences in forum threads
Sentence Parser
Brand word clusters containing "samsung", "htc", ...

229	hitec	rovio	samsun	ssg
503	klipsch	sam	**samsung**	sumsung
andino	magpul	sammy	samsungs	sung
aston	msgtypes	samseng	samung	vivendi
cnc	msung	samsuck	seagate	wright
fujifilm	netgear	samsumg	semaphore	

Model word clusters containing "galaxy", "active" ...
Brand Filter (4 Rules)
Brand variations

ssg, samseng, sam, **samsumg**, sammy, sumsung, samsun, sung, samsuck, samsung, samsungs, samung

samsung galaxy s3
samsung firmware
service center
Samsung updates...

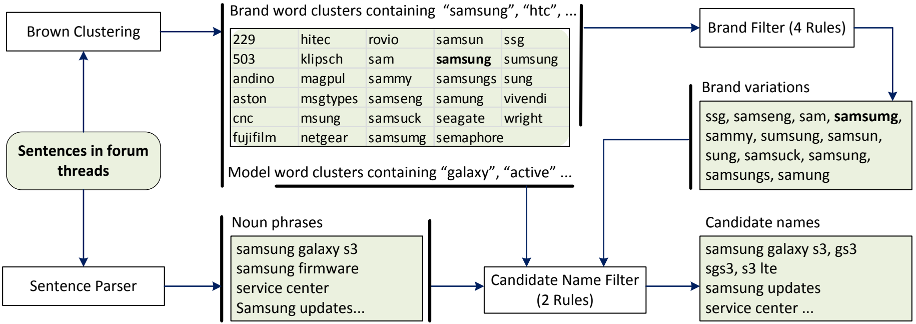
Candidate Name Filter (2 Rules)
Candidate names

samsung galaxy s3, gs3
sgs3, s3 lte
samsung updates
service center ...

Word cluster $\mathcal{W}_b$ contains brand *b*. A word $w \in \mathcal{W}_b$ is a variation of *b* if:

- The phonemic edit distance between *w* and *b* is 0, or
- The first and the last characters in *w* and *b* are the same, or
- The first three characters in *w* and *b* are the same, or
- Brand *b* contains more than one upper-case character and the prefix of *w* matches all upper-case characters in *b* in sequence (*e.g.*, bberry).

Candidate name generation


Brown Clustering
Sentences in forum threads
Sentence Parser
Brand word clusters containing "samsung", "htc", ...
Model word clusters containing "galaxy", "active" ...
Brand Filter (4 Rules)
Brand variations
Noun phrases
Candidate Name Filter (2 Rules)
Candidate names

Brand word clusters containing "samsung", "htc", ...

229	hitec	rovio	samsun	ssg
503	klipsch	sam	**samsung**	sumsung
andino	magpul	sammy	samsungs	sung
aston	msgtypes	samseng	samung	vivendi
cnc	msung	samsuck	seagate	wright
fujifilm	netgear	samsumg	semaphore	

Brand variations

ssg, samseng, sam, **samsumg**, sammy, sumsung, samsun, sung, samsuck, samsung, samsungs, samung

Noun phrases

samsung galaxy s3
samsung firmware
service center
Samsung updates...

Candidate names

samsung galaxy s3, gs3
sgs3, s3 lte
samsung updates
service center ...

A model word cluster contains at least one word in a phone model. A noun phrase is a candidate mobile phone name if it satisfies both rules:

1. The phrase contains a brand variation, or the phrase appears after a brand variation at least once in the whole dataset; and
2. At least one word in the phrase appears in a model word cluster and all the remaining words appear in either model word clusters or brand word clusters.

Name recognition: Conditional Random Field

Lexical features, **G**rammatical features, and **N**ame features

L_1	The current word and its surrounding two words $w_{i-2}w_{i-1}w_iw_{i+1}w_{i+2}$, and their lower-cased forms.
L_2	Word surface feature of the current word: Initial capitalization, all capitalization, containing capitalized letters, all digits, containing digits and letters.
G_1	POS tagger of the current word and its surrounding two words.
G_2	Path prefixes of length 4, 6, 10, 20 (*i.e.,* maximum length) of the current word by Brown clustering.
N_1	Flags to indicate whether the current word and its surrounding two words are candidate phone names
N_2	The brand entropy of the current word and its surrounding two words.

The key to build the CRF model: training data?

Names as queries for **automatic sentence labeling**

- **Positive names** $\mathcal{P}$:
  1. All formal names given as the input
  2. Formal names by replacing Roman number with Arabic number
  3. Model names if containing more than one word *e.g.,* “galaxy note”

- **Negative names** $\mathcal{N}$:
  Manual annotation from the set of candidate names $\mathcal{C}$, *e.g.,* “service center”, “firmware”, “update”**.**

- **33,072 sentences selected automatically**:
  1. The sentence contains at least one entity in either set $\mathcal{P}$ or set $\mathcal{N}$;
  2. The sentence does not contain any entry appears in $\mathcal{C} \setminus (\mathcal{P} \cup \mathcal{N})$

- **Candidate name as “*a single token*”**:
  Original sentence: Still prefer ip 5 then note 2
  Rewritten sentence: Still prefer ip_5 then note_2

Name normalization: Lexical rules

Most phone name variations detected are originated from the candidate name set $\mathcal{C} \rightarrow$ Candidate names in $\mathcal{C}$ can be pre-normalized.

Normalization

- Sequence containment
  “SGS III” are contained in “Samsung Galaxy SIII”
- Model number containment
  “i9300”, “i9305”, “s3 i9300”, and “samsung i9300 galaxy s iii”
- Confidence score
  Number of appearances in threads titled with formal names. “SGS II” matches “Samsung Galaxy SII” and “Samsung Galaxy SIII”.

Experiments: Data and ground truth

Forum data:

- HardwareZone forum: “Mobile Communication Technology”. 1,026,190 posts in 25,251 threads from March 2002 to May 2013.
- Formal names from GSMArena.com

Ground truth labelling:

- 20 most popular phones of 8 brands, one thread per phone
- 4,121 sentences with 946 phone name mentions.

Apple, HTC, LG	–No brand variations–
Nokia	nokia, nokie, nk
BlackBerry	blackberry, bbry, blackbery, bb, bberry
Motorola	motorola, moto, motorolla, mot
Samsung	ssg, samseng, sam, samsumg, sammy, sumsung, samsun, sung, samsuck, samsung, samsungs, samung
Sony Erricson	sony erricson, sony ericsson, sony ericson, sony ericcson, sonyericsson, sony ericssion, sn, sony, sonyeric

Experiments: Methods and evaluation metric

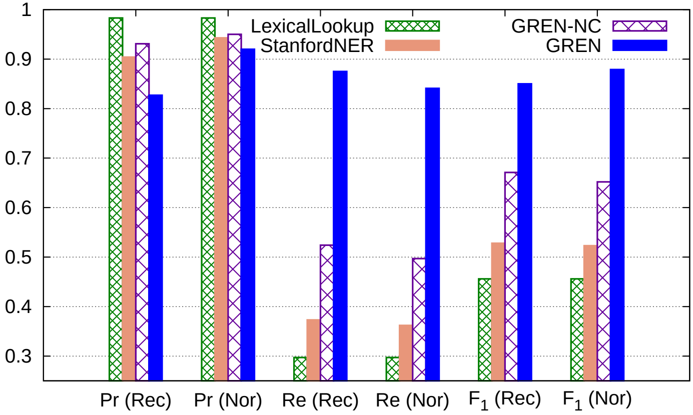
Method comparison: trained on the same set of 33,072 sentences.

1. **LexicalLookup.** Formal names used as a dictionary.
2. **StanfordNER.** Use default features provided by the package
3. **GREN**: The proposed method with candidate name generation, CRF-based name recognition, and rule-based normalization.
4. **GREN-NC.** Use the same set of features as GREN but not re-writing the sentences.

Evaluation metric

- **Precision (Pr)**: the ratio of true phone name mentions among all mentions that are predicted positively.
- **Recall (Re)**: the ratio of correctly recognized name mentions among all phone name mentions annotated in the ground truth data.
- F_1: the harmonic mean of precision and recall.

Experimental results



Rec: name recognition Nor: name normalization

Summary

Lessons learnt:

1. Brown clustering is effective in “grouping” product name variants
2. Rule-based approach is useful in product name recognition if there exist naming convention
3. Large number of training examples is necessary for effective NER
4. With rule-based approach, training examples can be obtained in semi-automatic manner

Limitations:

1. Candidate name set needs to be updated from time to time
2. Code name cannot be normalized to phone name
   e.g., “Nozomi” to “Sony Xperia S”

NANYANG TECHNOLOGICAL UNIVERSITY

Dr. Aixin SUN
axsun@ntu.edu.sg
http://www.ntu.edu.sg/home/axsun/