中国科学院新疆理化技术研究所
XINJIANG TECHNICAL INSTITUTE OF PHYSICS AND CHEMISTRY, CAS

Detection of Loan Words in Uyghur Texts

- Chenggang Mi, Yating Yang, Lei Wang, Xiao Li and Kamali Dalielihan
- **Xinjiang Technical Institute of Physics & Chemistry of**
- **Chinese Academy of Sciences**
- michenggang@gmail.com,{yangyt,wanglei,xiaoli}@ms.xjb.ac.cn,
- kamaly330@gmail.com

Outline

- Motivation
- Our method
- Experiments and analysis
- Conclusion and future work

Motivation

- For tasks like SMT, there always exist data sparseness during training of translation models because lack of bilingual texts.
- There are many loan words in Uyghur, which are mainly borrowed from Chinese and Russian.

Motivation

Chinese loan words in Uyghur [in English]		Russian loan words in Uyghur [in English]	
شىنجاڭ(新疆)	[Xinjiang]	رومكا(рюмка)	[cup]
لەگمەن(拉面)	[noodles]	تېلېفون(телефон)	[telephone]
لازا(辣子)	[hot pepper]	ئۇنىۋېرسىتېت(университет)	[university]
شۇجى(书记)	[secretary]	رادىيو(радио)	[radio]
كوي(块)	[Yuan]	پوچتا(почта)	[post office]
لەگپۇلا(凉粉)	[agar-agar jelly]	ۋېلسىپىت(велосипед)	[bicycle]
دوفۇ(豆腐)	[bean curd]	ئوبلاست(область)	[region]

Motivation

- A loanword is a word borrowed from a don or language and incorporated into a recipi ent language directly, without translation.
- To enrich bilingual resources, we detect Chinese and Russian loan words from Uyg hur texts according to phonetic similarities between a loan word and its correspondin g donor language word.

Our Method

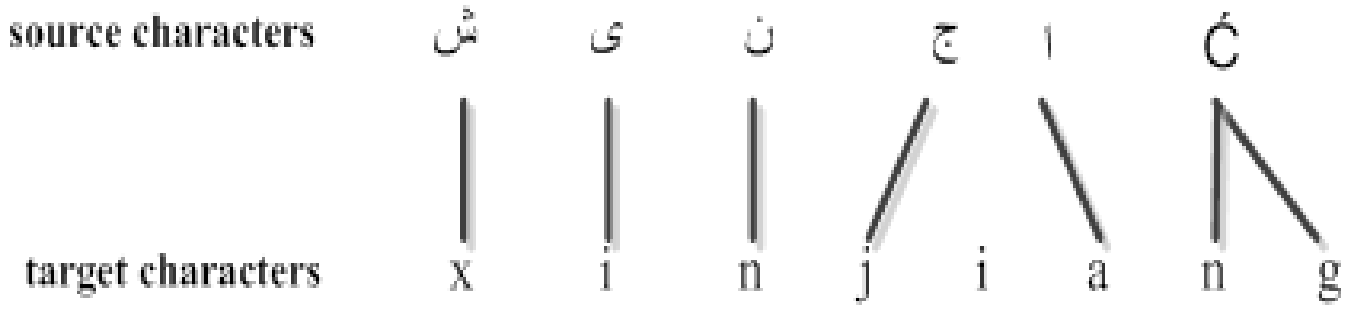
- We consider the loan words detection as a binary classification problem.
- Features used during the model training are computed by five string similarity algorithms.
- We obtain character mapping rules by character aligning.

Our Method

- Character Alignment

source: Uyghur character sequence

target: Chinese Pinyin (Russian Latin) character sequence



Our Method

- Character Alignment

source: Uyghur character sequence

target: Chinese Pinyin (Russian Latin) character sequence

1) IBM model 1

$$p(e,a \mid f) = \frac{\varepsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t(e_j \mid f_{a(j)}) \quad , 1 \le j \le l_e .$$

2) HMM

$$P(f \mid e) = \sum_{a_1^m} \prod_{j=1}^{m} \left[p(a_j \mid a_{j-1}, l) \bullet p(f_j \mid e_{a_j}) \right]$$

Our Method

- Position-Related Edit Distance

$$ED_{a,b}(i,j)=\begin{cases}\max(i,j) & if\ \min(i,j)=0,\\ \min\begin{cases}ED_{a,b}(i-1,j)+1\\ ED_{a,b}(i,j-1)+1\\ ED_{a,b}(i-1,j-1)+1_{(ai\neq bj)}\end{cases} & otherwise.\end{cases}$$

$$PRED_{a,b}(i,j)=\begin{cases}ED_{a,b}(i,j) & no\ continue\ delete\ occurred,\\ ED_{a,b}(i,j)-times_{delet\_occur\_end}(a,b) & continue\ delete\ occurred\end{cases}$$

Our Method

- Weighted Common Subsequence

We assign a weight to each common subsequence according to its length.

$$WCS_{a,b} = \sum_{i=2}^{\min(La,Lb)} LEN_i \bullet NUM_i$$

Our Method

- Perceptron-Based Loan Words Detection

We consider detection of loan words in Uyghur texts as a perceptron-based classification problem.

$$f(x) = \begin{cases} 1 & if \quad w \bullet x + b > 0 \\ 0 & otherwise \end{cases}$$

x: $< PRED_{a,b}, DC_{a,b}, WCS_{a,b}, JSC_{a,b}, OLS_{a,b} >$

Experiments and Analysis

- Data sets

  -Character aligment

  Uyghur-Chinese entities corpus: 200 word pairs

  Uyghur-Russian entities corpus: 200 word pairs

  -Loan words dectection

  1,000 Uyghur sentences (selected from web)

Experiments and Analysis

- Character Alignment

Aligning of the source character sequence and the target character sequence is just as a word alignment procedure.

- Dectection of loan words in Uyghur texts

Features of each word pair and its loan word label will be considered as an input of the perceptron-based detection model.

Experiments and Analysis

	Chinese Loan Words			Russian Loan Words		
	R	P	F1	R	P	F1
ED	71.20	62.41	66.52	73.29	67.32	70.18
DC	69.22	60.98	64.84	70.02	62.41	66.00
CS	73.12	61.16	66.61	76.08	68.43	72.05
OS	69.25	60.73	64.71	70.29	63.78	66.88
JSC	69.10	61.98	65.35	71.81	64.59	68.01

ED: Edit Distance, DC: Dice Coefficient, CS: Common Subsequence, OS: Overlap Similarity, JSC: Jaccard Similarity Coefficient

Table 1 Evaluation of five basic algorithms

Experiments and Analysis

	Chinese Loan Words			Russian Loan Words		
	R	P	F1	R	P	F1
PRED	75.72	64.73	69.80	75.39	70.02	72.61
DC	69.78	62.33	66.35	71.64	63.25	67.18
WCS	74.39	64.36	69.01	78.01	72.34	75.07
OS	71.29	61.72	66.16	71.05	65.20	68.00
JSC	71.32	63.65	67.27	72.89	65.37	68.92

PRED: Position-Related Edit Distance, DC: Dice Coefficient, WCS: Weighted Common Subsequence, OS: Overlap Similarity, JSC: Jaccard Similarity Coefficient

Table 2 Evaluation of five algorithms (ED->PRED, CS->WCS)

Experiments and Analysis

	Chinese Loan Words			Russian Loan Words		
	R	P	F1	R	P	F1
PBDM	78.82	68.30	73.18	81.03	73.22	76.93

PBDM: Perceptron-Based Detection Model

Table 3 Perceptron-based loan words detection model

Experiments and Analysis

- Performance of PRED and WCS outperform ED and CS, significantly.
- Results of Table 2 are based on mapping rules obtained by character aligning, which contribute to the performance of string similarity algorithms.

Experiments and Analysis

- Among our experiments, PBDM achieved the best performance.

  - PBDM integrate advantage of five string similarity algorithms.

  - The error-driven model much adaptive to our task.

- Performance of Russian loan words detection outperform Chinese.

  - The spelling style of Russian loan words is much closer with Uyghur.

Conclusion

- We transfer the phonetic similarity between donor language words and Uyghur words to string similarity, and consider the detecting procedure as a classification problem.
- Experimental results show that with our model, Chinese and Russian loan words can be recognized efficiently.

Future work

In future work, we will focus on:

- extraction of bilingual resources based on loan words;
- extend our model to other languages.

Thanks