Identification of Transliterated Foreign Words in Hebrew Script

Yoav Goldberg Michael Elhadad

Ben-Gurion University
of the Negev

CiCLing 2008, Haifa, Israel

A Typical Hebrew Text

Taken from YNET Gossip Section a Few Days Ago

gossip רכילות ובידור

אולי הפעם נתעלס?

סקס, אהבה וסמו

אורי פפר, תאכל אבק! בקרב הבנות בקאסט של "אולי הפעם" יש תמימות דעים: את הוולנטיינ'ז דיי הכי טוב לבלות עם יובל סמו. וכולן יודעות גם מה יעשו איתו

קרן נתנזון
פורסם: 14.02.08, 18:02

A Typical Hebrew Text

Taken from YNET Gossip Section a Few Days Ago

תרבות ובידור › רכילות › כחול לבן

gossip רכילות ובידור

ריחול • כחול לבן • הגולש האמיץ • דאבל דא • על הסט • העצומה • מרחק נגיעה

סקס, אהבה וסמו

אולי הפעם נתעלס?

אורי פפר, תאכל אבק! בקרב הבנות בקאסט של "אולי הפעם" יש תמימות דעים: את הוולנטיינ'ז דיי הכי טוב לבלות עם יובל סמו. וכולן יודעות גם מה יעשו איתו

קרן נתנזון
פורסם: 14.02.08, 18:02

A Typical Hebrew Text

Taken from YNET Gossip Section a Few Days Ago

KAST VVLLNTYYN'Z DYY ST DABL SKS

A Typical Hebrew Text

Taken from YNET Gossip Section a Few Days Ago

תרבות ובידור › רכילות › כחול לבן

gossip רכילות ובידור

ריחול • כחול לבן • הגולש האמיץ • דאבל דא • על הסט • העצומה • מרחק נגיעה

אולי הפעם נתעלס?

אורי פפר, תאכל אבק! בקרב הבנות בקאסט של "אולי הפעם" יש תמימות דעים: את הוולנטיינ'ז דיי הכי טוב לבלות עם יובל סמו. וכולן יודעות גם מה יעשו איתו

קרן נתנזון
פורסם: 14.02.08, 18:02

סקס, אהבה וסמו

- Foreign words written in Hebrew script
- Can't expect comprehensive dictionary coverage
- Would like to identify them automatically

Which words are we looking for?

Names

- of people if they are not Israeli / Hebrew / Russian / Amharic / Arabic
- of places in case they are pronounced the same in English
- of Companies/Organization if they sound non-Hebrew (mostly easy to decide)
- of Months if they sound the same in English

Example

- YES: ג'ון/John, רוברט/Robert, ג'ייק/Jake
- NO: יואב/Yoav, גולדברג/Goldberg, אלחַדד/Elhadad

Which words are we looking for?

Names

- of people if they are not Israeli / Hebrew / Russian / Amharic / Arabic
- of places in case they are pronounced the same in English
- of Companies/Organization if they sound non-Hebrew (mostly easy to decide)
- of Months if they sound the same in English

Example

- YES: קרן/Karen
- NO: קרן/Keren

Which words are we looking for?

Names

- of people if they are not Israeli / Hebrew / Russian / Amharic / Arabic
- of places in case they are pronounced the same in English
- of Companies/Organization if they sound non-Hebrew (mostly easy to decide)
- of Months if they sound the same in English

Example

- YES: מייקל/Michael (pronounced maykel)
- NO: מיכאל/Michael (pronounced mi-cha-el)

Which words are we looking for?

Names

- of people if they are not Israeli / Hebrew / Russian / Amharic / Arabic
- of places in case they are pronounced the same in English
- of Companies/Organization if they sound non-Hebrew (mostly easy to decide)
- of Months if they sound the same in English

Example

- YES: הוליווד/Hollywood, ניו יורק/New-York
- NO: אנגליה/Anglia (England), איטליה/Italya (Italy)

Which words are we looking for?

Names

- of people if they are not Israeli / Hebrew / Russian / Amharic / Arabic
- of places in case they are pronounced the same in English
- of Companies/Organization if they sound non-Hebrew (mostly easy to decide)
- of Months if they sound the same in English

Example

- YES: מייקרוסופט/Microsoft, נייק/Nike
- NO: אוסם/Osem, תנובה/Tnuva

Which words are we looking for?

Names

- of people if they are not Israeli / Hebrew / Russian / Amharic / Arabic
- of places in case they are pronounced the same in English
- of Companies/Organization if they sound non-Hebrew (mostly easy to decide)
- of Months if they sound the same in English

Example

- YES: אוגוסט/August, ספטמבר/September
- NO: יוני/Yuni (June), יולי/Yuli (July)

Which words are we looking for?

Cognates, transliterations and borrowings

It revolves around the pronunciation

- Some words are clearly Foreign
- Foreign origin, but Hebrew-sounding – NO
- Non inflected, and pronounced the same – YES
- Inflected, pronounced the same – YES
- Inflected, pronounced differently – MAYBE
- Can be read as Hebrew or Foreign – DEPENDS on context

Example

קאסט/Cast, דאבל/Double, דיי/Day, טרנד/Trend

Which words are we looking for?

Cognates, transliterations and borrowings

It revolves around the pronunciation

- Some words are clearly Foreign
- Foreign origin, but Hebrew-sounding – NO
- Non inflected, and pronounced the same – YES
- Inflected, pronounced the same – YES
- Inflected, pronounced differently – MAYBE
- Can be read as Hebrew or Foreign – DEPENDS on context

Example

אנציקלופדיה/En-ci-klo-pe-di-ya

Which words are we looking for?

Cognates, transliterations and borrowings

It revolves around the pronunciation

- Some words are clearly Foreign
- Foreign origin, but Hebrew-sounding – NO
- Non inflected, and pronounced the same – YES
- Inflected, pronounced the same – YES
- Inflected, pronounced differently – MAYBE
- Can be read as Hebrew or Foreign – DEPENDS on context

Example

רדיו/Radio, סקס/Sex

Which words are we looking for?

Cognates, transliterations and borrowings

It revolves around the pronunciation

- Some words are clearly Foreign
- Foreign origin, but Hebrew-sounding – NO
- Non inflected, and pronounced the same – YES
- Inflected, pronounced the same – YES
- Inflected, pronounced differently – MAYBE
- Can be read as Hebrew or Foreign – DEPENDS on context

Example

טרנדי/Trendy

Which words are we looking for?

Cognates, transliterations and borrowings

It revolves around the pronunciation

- Some words are clearly Foreign
- Foreign origin, but Hebrew-sounding – NO
- Non inflected, and pronounced the same – YES
- Inflected, pronounced the same – YES
- Inflected, pronounced differently – MAYBE
- Can be read as Hebrew or Foreign – DEPENDS on context

Example

אלכוהולי/Alcoholi (vs. Alcoholic)

Which words are we looking for?

Cognates, transliterations and borrowings

It revolves around the pronunciation

- Some words are clearly Foreign
- Foreign origin, but Hebrew-sounding – NO
- Non inflected, and pronounced the same – YES
- Inflected, pronounced the same – YES
- Inflected, pronounced differently – MAYBE
- Can be read as Hebrew or Foreign – DEPENDS on context

Example

בד/Bad vs. *cloth*, *branch*, רן/Run vs. *sang*,Proper Name

The approach

We chose to tackle the problem as

Performing
Language Identification
at the
Word Level

The approach

We chose to tackle the problem as

Performing
Language Identification
at the
Word Level

- Language Identification accuracy: $> 99\%$
- A solved problem?

The approach

We chose to tackle the problem as

Performing
Language Identification
at the
Word Level

- Language Identification accuracy: $> 99\%$
- . . . but requires about 50 characters

The approach

We chose to tackle the problem as

Performing
Language Identification
at the
Word Level

- Language Identification accuracy: $> 99\%$
- . . . but requires about 50 characters
- . . . and on top of that. . .

Additional Problems We Have to Face

Hebrew Writing System

- Vowels are not written in most cases
- Letters (א,ו,י) can be either vowels or consonants
- Each of (*p*,*f*) (*b*,*v*) (*s*,*sh*) are encoded by the same letter (פ,ב,ש)
- Sounds *th*, *j* do not exist

$\Rightarrow$ Words are even shorter

$\Rightarrow$ Words forms are very ambiguous

Additional Problems We Have to Face

Hebrew Writing System

- Vowels are not written in most cases
- Letters (א,ו,י) can be either vowels or consonants
- Each of (*p*,*f*) (*b*,*v*) (*s*,*sh*) are encoded by the same letter (פ,ב,ש)
- Sounds *th*, *j* do not exist

⇒ Words are even shorter

⇒ Words forms are very ambiguous

(Lack of) Training Data

- No dictionary available
- Many ways for transliterating the same word

Naive Bayes Language Identification

English

- Assume a language model M generating a word w
- We have several such models (M_i), one for each language.
- For a given word, we would like to find the language L most likely to generate this word
- Using Bayes rule , the word is fixed , assume models are a-priori equally likely

Math

$$p(w|M)$$

Naive Bayes Language Identification

English

- Assume a language model M generating a word w
- We have several such models (M_i), one for each language.
- For a given word, we would like to find the language L most likely to generate this word
- Using Bayes rule , the word is fixed , assume models are a-priori equally likely

Math

$$L = \arg\max_{L} p(M_L|w)$$

Naive Bayes Language Identification

English

- Assume a language model M generating a word w
- We have several such models (M_i), one for each language.
- For a given word, we would like to find the language L most likely to generate this word
- Using Bayes rule , the word is fixed , assume models are a-priori equally likely

Math

$$P(M_i|w) = P(w|M_i)P(M_i)/P(w)$$

Naive Bayes Language Identification

English

- Assume a language model M generating a word w
- We have several such models (M_i), one for each language.
- For a given word, we would like to find the language L most likely to generate this word
- Using Bayes rule , the word is fixed , assume models are a-priori equally likely

Math

$$P(M_i|w) \propto P(w|M_i)P(M_i)$$

Naive Bayes Language Identification

English

- Assume a language model M generating a word w
- We have several such models (M_i), one for each language.
- For a given word, we would like to find the language L most likely to generate this word
- Using Bayes rule , the word is fixed , assume models are a-priori equally likely

Math

$$P(M_i|w) \propto P(w|M_i)$$

Naive Bayes Language Identification

How do we estimate

$$P(w|M_i)$$

?

Probabilistic Model
Generative N-Gram Model

English

- A Markov Model generating letters
- Each letter depends on the previous n letters

Math

Probability for generating a letter (c):

$$P(c_i|c_{i-1}, \ldots, c_{i-n})$$

Probabilistic Model
Generative N-Gram Model

English

- A Markov Model generating letters
- Each letter depends on the previous n letters

Math

Probability for generating a word ($w = c_1, \ldots, c_N$):

$$\prod_{n \leq i \leq N} P(c_i | w_{i-1}, \ldots, c_{i-n})$$

Probabilistic Model
Generative N-Gram Model

English

- A Markov Model generating letters
- Each letter depends on the previous n letters
- The probabilities for generating a letter are the Model Parameters

Math

Model Parameters are estimated via Relative Frequency:

$$P(c_i|c_{i-1},\ldots,c_{i-n}) = \frac{\#(c_i c_{i-1}\ldots c_{i-n})}{\#(c_{i-1}\ldots c_{i-n})}$$

Probabilistic Model
Non-Traditional Smoothing

English

- Smoothing is hardly needed for 1- or 2-gram models.
- . . . But we want the power of 3- and 4-gram models.

Math

Probabilistic Model
Non-Traditional Smoothing

English

- Smoothing is hardly needed for 1- or 2-gram models.
- . . . But we want the power of 3- and 4-gram models.
- Instead of using a traditional backoff strategy, we linearly combine 4 different n-gram models

Math

$$P(w|M_i) = \lambda_1 P(w|M_{i1}) + \lambda_2 P(w|M_{i2}) + \lambda_3 P(w|M_{i3}) + \lambda_4 P(w|M_{i4})$$

Probabilistic Model
Non-Traditional Smoothing

English

- Smoothing is hardly needed for 1- or 2-gram models.
- . . . But we want the power of 3- and 4-gram models.
- Instead of using a traditional backoff strategy, we linearly combine 4 different n-gram models (this can be viewed as a *voting scheme*)

Math

$$P(w|M_i) = \lambda_1 P(w|M_{i1}) + \lambda_2 P(w|M_{i2}) + \lambda_3 P(w|M_{i3}) + \lambda_4 P(w|M_{i4})$$

Probabilistic Model
Non-Traditional Smoothing

English

- Smoothing is hardly needed for 1- or 2-gram models.
- . . . But we want the power of 3- and 4-gram models.
- Instead of using a traditional backoff strategy, we linearly combine 4 different n-gram models (this can be viewed as a *voting scheme*)
- We didn't do anything fancy for setting the λs

Math

$$P(w|M_i) = \lambda_1 P(w|M_{i1}) + \lambda_2 P(w|M_{i2}) + \lambda_3 P(w|M_{i3}) + \lambda_4 P(w|M_{i4})$$

Probabilistic Model
Backward Model

English

- We also add a backward moving models
- (some of the probabilities are different. . .)

Probabilistic Model
Backward Model

English

- We also add a backward moving models
- (some of the probabilities are different...)

Math

$$P(w|M_i) = \lambda_1 P(w|M_{i1}) + \lambda_2 P(w|M_{i2}) + \lambda_3 P(w|M_{i3}) + \lambda_4 P(w|M_{i4}) + \lambda_5 P(w|M_{i2back}) + \lambda_6 P(w|M_{i3back}) + \lambda_7 P(w|M_{i4back})$$

Unsupervised Setting: Training Data by Over-generation

Problem

- We need to count the number of times each ngram occurs
- . . . But we don't have transliterated-Hebrew text
- . . . And we don't even have "pure Hebrew" text

Unsupervised Setting: Training Data by Over-generation

Problem

- We need to count the number of times each ngram occurs
- . . . But we don't have transliterated-Hebrew text
- **. . . And we don't even have "pure Hebrew" text**

Solution 1

Using `Ben Yehuda Project` for
estimating *pure Hebrew*

Unsupervised Setting: Training Data by Over-generation

Problem

- We need to count the number of times each ngram occurs
- **. . . But we don't have transliterated-Hebrew text**
- . . . And we don't even have "pure Hebrew" text

Solution 2

Using the `Brown Corpus`,
the `CMU pronunciation dictionary`
and `a few simple rules` for
generating *noisy transliterated-Hebrew*

The Over-Generation Process

Raw English Text 6mb of Brown

```
The Fulton County Grand
Jury said Friday an
investigation of
Atlanta's recent ...
```

The Over-Generation Process

Raw English Text 6mb of Brown

```
The Fulton County Grand
Jury said Friday an
investigation of
Atlanta's recent ...
```

The Over-Generation Process

Raw English Text 6mb of Brown

```
The Fulton County Grand
Jury said Friday an
investigation of
Atlanta's recent ...
```

Phonetic Representation

- *K AW N T IY
- *K AW N IY

The Over-Generation Process

Raw English Text 6mb of Brown

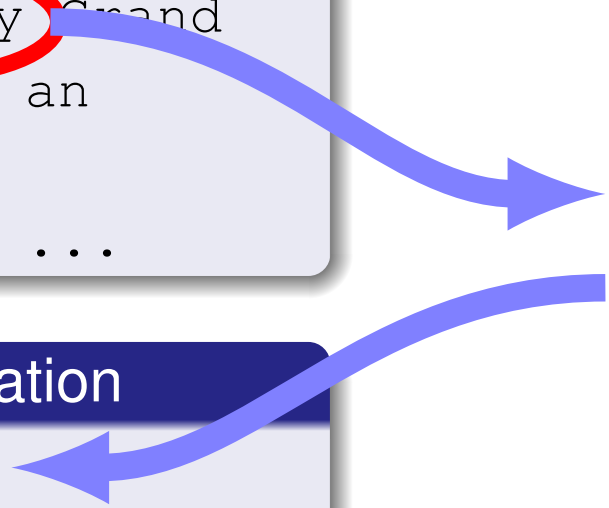
The Fulton County Grand
Jury said Friday an
investigation of
Atlanta's recent ...

Phonetic Representation

- *K AW N T IY
- *K AW N IY

CMUDICT

...
COUNTS K AW1 N T S
COUNTY K AW1 N T IY0
COUNTY-2 K AW1 N IY0
COUP K UW1
...



The Over-Generation Process

Raw English Text 6mb of Brown

```
The Fulton County Grand
Jury said Friday an
investigation of
Atlanta's recent ...
```

Phonetic Representation

- *K AW N T IY
- *K AW N IY

CMUDICT

```
...
COUNTS K AW1 N T S
COUNTY K AW1 N T IY0
COUNTY-2 K AW1 N IY0
COUP K UW1
...
```

Phonetic → Hebrew

*K	כ
AW	ואו, אוו, או
N	נ
T	ת, ט
IY	י, אי, ϵ

The Over-Generation Process

Raw English Text 6mb of Brown

```
The Fulton County Grand
Jury said Friday an
investigation of
Atlanta's recent ...
```

Phonetic Representation

- *K AW N T IY
- *K AW N IY

Many Hebrew Transliterations

קאונתי קאונתאי קאונת **קאונטי** קאונטאי
קאונט קאוונתי קאוונתאי קאוונת **קאוונטי**
קאוונטאי קאוונט קואונתי קואונתאי קואונת
קואונטי קואונטאי קואונט קאוני קאונאי קאונ
קאווני קאוונאי קאוונ קואוני קואונאי קואונ

CMUDICT

```
...
COUNTS K AW1 N T S
COUNTY K AW1 N T IY0
COUNTY-2 K AW1 N IY0
COUP K UW1
...
```

Phonetic → Hebrew

*K	כ
AW	ואו, אוו, או
N	נ
T	ת, ט
IY	י, אי, ϵ

More about Over-Generation

Isn't This Just Like Writing Rules?

- Not really

More about Over-Generation

Isn't This Just Like Writing Rules?

- Not really
- Simple Rules $\rightarrow$ Data $\rightarrow$ Learning $\rightarrow$ Complex Rules of a Different Nature

More about Over-Generation

Isn't This Just Like Writing Rules?

- Not really
- Simple Rules $\rightarrow$ Data $\rightarrow$ Learning $\rightarrow$ Complex Rules of a Different Nature

Some Things to Notice

- Every foreign word is represented in accordance to **actual corpus frequency**
- Many writing variations are taken into account

$\Rightarrow$ Complex interactions

Not All Foreign Words are Created Equal

Observation

- Many of the transliterated words are Proper Names
- The sound patterns of proper names are somewhat different than those of regular words

Not All Foreign Words are Created Equal

Observation

- Many of the transliterated words are Proper Names
- The sound patterns of proper names are somewhat different than those of regular words

Adding another language model

⇒ We actually have 3 languages:
Hebrew, Foreign and Foreign-Name

- Use a heuristic for extracting proper nouns, and estimate a language model for it
- This is very noisy – **but still useful**
  If the classifier decides Foreign **or** Foreign-Name, we take it to be Foreign

Adding a Lexicon

HSpell – a Hebrew Spell Checker

- Lets assume a word is Foreign IFF its not in HSpell

$\Rightarrow$ Not very good results

- Lets assume a word is Foreign IFF its not in HSPELL and the statistical model says its Foreign

$\Rightarrow$ Better

Adding a Lexicon

HSpell – a Hebrew Spell Checker

- Lets assume a word is Foreign IFF its not in HSpell

$\Rightarrow$ Not very good results

- Lets assume a word is Foreign IFF its not in HSPELL and the statistical model says its Foreign

$\Rightarrow$ Better

Evaluation Setting

The Data

- 50 articles from gossip section of YNET
- 9618 words, 4044 word types
- Removed prefixes
- **About 10% of the words are Foreign!**
- 3608 Hebrew words
- 251 Foreign Proper Names
- 117 Foreign Words
- 68 Ambiguous (hard for Humans to judge) – discarded

Results - Supervised

5-fold cross validation experiment

Experiment	Precision (%)	Recall (%)
Baseline	17.75	60.42
Vot	59.72	60.81
Vot+Back	59.70	60.76

Results - Supervised

5-fold cross validation experiment

Experiment	Precision (%)	Recall (%)
Baseline	17.75	60.42
Vot	59.72	60.81
Vot+Back	59.70	60.76

Baseline: state of the art Language Identification (Dunnning 94)

Results - Supervised

5-fold cross validation experiment

Experiment	Precision (%)	Recall (%)
Baseline	17.75	60.42
Vot	59.72	60.81
Vot+Back	59.70	60.76

Our combined (smoothed) models, with and without backward models

Results – **Un**supervised

Training on Ben-Yehuda and Generated Data

Experiment	Precision (%)	Recall (%)
Best Supervised	59.70	60.76
Basline(3gram)	58.7	64.9
Basline(4gram)	55.6	65.7
Vot	76.3	71.7
Vot+Back	80.4	71.4
Vot+Back+Names	80.1	82

Results – **Un**supervised

Training on Ben-Yehuda and Generated Data

Experiment	Precision (%)	Recall (%)
Best Supervised	59.70	60.76
Basline(3gram)	58.7	64.9
Basline(4gram)	55.6	65.7
Vot	76.3	71.7
Vot+Back	80.4	71.4
Vot+Back+Names	80.1	82

Baseline: state-of-the-art language identification
3- and 4-gram models

Results – **Un**supervised

Training on Ben-Yehuda and Generated Data

Experiment	Precision (%)	Recall (%)
Best Supervised	59.70	60.76
Basline(3gram)	58.7	64.9
Basline(4gram)	55.6	65.7
Vot	76.3	71.7
Vot+Back	80.4	71.4
Vot+Back+Names	80.1	82

Our model: smoothed

Results – **Un**supervised

Training on Ben-Yehuda and Generated Data

Experiment	Precision (%)	Recall (%)
Best Supervised	59.70	60.76
Basline(3gram)	58.7	64.9
Basline(4gram)	55.6	65.7
Vot	76.3	71.7
Vot+Back	80.4	71.4
Vot+Back+Names	80.1	82

Our model: adding a backward model

Results – **Un**supervised

Training on Ben-Yehuda and Generated Data

Experiment	Precision (%)	Recall (%)
Best Supervised	59.70	60.76
Basline(3gram)	58.7	64.9
Basline(4gram)	55.6	65.7
Vot	76.3	71.7
Vot+Back	80.4	71.4
Vot+Back+Names	80.1	82

Our model: adding a proper names model

Results – **Un**supervised with a Lexicon

Results Using HSpell

Experiment	Precision (%)	Recall (%)
Best w/o lexicon	80.1	82
HSPELL	13	85
Vot+Back+HSPELL	91.86	61.41
Vot+Back+Names+HSPELL	91.19	70.38

Results – **Un**supervised with a Lexicon

Results Using HSpell

Experiment	Precision (%)	Recall (%)
Best w/o lexicon	80.1	82
HSPELL	13	85
Vot+Back+HSPELL	91.86	61.41
Vot+Back+Names+HSPELL	91.19	70.38

Only HSpell

Results – **Un**supervised with a Lexicon

Results Using HSpell

Experiment	Precision (%)	Recall (%)
Best w/o lexicon	80.1	82
HSPELL	13	85
Vot+Back+HSPELL	91.86	61.41
Vot+Back+Names+HSPELL	91.19	70.38

HSPell + Statistical Model (no Proper Names model)

Results – **Un**supervised with a Lexicon

Results Using HSpell

Experiment	Precision (%)	Recall (%)
Best w/o lexicon	80.1	82
HSPELL	13	85
Vot+Back+HSPELL	91.86	61.41
Vot+Back+Names+HSPELL	91.19	70.38

HSPELL + Statistical Model (with Proper Names model)

Conclusions and Summary

The Good

- We identify Foreign Words in Hebrew Script
- We do it quite accurately . . .
- . . . Without annotating Hebrew data
- Technique is language independent

The Interesting

- Lots of noisy data is WAY better than little clean data

The Bad

- Completely context free – some decisions are impossible
- We don't treat משהוכלב prefixes
- We didn't actually try any other language pair

Thanks