Einführung in die Computerlinguistik
Word Shapes

Hinrich Schütze & Robert Zangenfeind

Centrum für Informations- und Sprachverarbeitung, LMU München

2015-10-16

How big is the term vocabulary?

- That is, how many types (distinct words) are there?
- Can we assume there is an upper bound?
- Not really: At least $70^{20} \approx 10^{37}$ different words of length 20.
- The vocabulary will keep growing with collection size.

Word shapes

Word shapes

- Replace all lower case letters with 'x'

Word shapes

- Replace all lower case letters with 'x'
- Replace all upper case letters with 'X'

Word shapes

- Replace all lower case letters with 'x'
- Replace all upper case letters with 'X'
- Replace all digits letters with '9'

Word shapes

- Replace all lower case letters with 'x'
- Replace all upper case letters with 'X'
- Replace all digits letters with '9'
- "Deduplicate": any sequence of $n > 1$ identical characters c is replaced by a single copy of c

Word shapes

- Replace all lower case letters with 'x'
- Replace all upper case letters with 'X'
- Replace all digits letters with '9'
- "Deduplicate": any sequence of $n > 1$ identical characters c is replaced by a single copy of c
- "submarine" -> "xxxxxxxxx" -> "x"

Word shapes

- Replace all lower case letters with 'x'
- Replace all upper case letters with 'X'
- Replace all digits letters with '9'
- "Deduplicate": any sequence of $n > 1$ identical characters c is replaced by a single copy of c
- "submarine" -> "xxxxxxxxx" -> "x"
- "London" -> "Xxxxxx" -> "Xx"

Word shapes

- Replace all lower case letters with 'x'
- Replace all upper case letters with 'X'
- Replace all digits letters with '9'
- "Deduplicate": any sequence of $n > 1$ identical characters c is replaced by a single copy of c
- "submarine" -> "xxxxxxxxx" -> "x"
- "London" -> "Xxxxxx" -> "Xx"
- "half-filled" -> "xxxx-xxxxxx" -> "Xx"

Word shapes

- Replace all lower case letters with 'x'
- Replace all upper case letters with 'X'
- Replace all digits letters with '9'
- "Deduplicate": any sequence of $n > 1$ identical characters c is replaced by a single copy of c
- "submarine" -> "xxxxxxxxx" -> "x"
- "London" -> "Xxxxxx" -> "Xx"
- "half-filled" -> "xxxx-xxxxxx" -> "Xx"
- "LMU" -> "XXX" -> "X"

Word shapes

- Replace all lower case letters with 'x'
- Replace all upper case letters with 'X'
- Replace all digits letters with '9'
- "Deduplicate": any sequence of $n > 1$ identical characters c is replaced by a single copy of c
- "submarine" -> "xxxxxxxxx" -> "x"
- "London" -> "Xxxxxx" -> "Xx"
- "half-filled" -> "xxxx-xxxxxx" -> "Xx"
- "LMU" -> "XXX" -> "X"
- "2015" -> "9999" -> "9"

Word shapes

- Replace all lower case letters with 'x'
- Replace all upper case letters with 'X'
- Replace all digits letters with '9'
- "Deduplicate": any sequence of $n > 1$ identical characters c is replaced by a single copy of c
- "submarine" -> "xxxxxxxxx" -> "x"
- "London" -> "Xxxxxx" -> "Xx"
- "half-filled" -> "xxxx-xxxxxx" -> "Xx"
- "LMU" -> "XXX" -> "X"
- "2015" -> "9999" -> "9"
- "3.1415" -> "9.9999" -> "9.9"

Word shapes

- Replace all lower case letters with 'x'
- Replace all upper case letters with 'X'
- Replace all digits letters with '9'
- "Deduplicate": any sequence of $n > 1$ identical characters c is replaced by a single copy of c
- "submarine" -> "xxxxxxxxx" -> "x"
- "London" -> "Xxxxxx" -> "Xx"
- "half-filled" -> "xxxx-xxxxxx" -> "Xx"
- "LMU" -> "XXX" -> "X"
- "2015" -> "9999" -> "9"
- "3.1415" -> "9.9999" -> "9.9"
- "Yahoo!" -> "Xxxxx!" -> "Xx!"

Frequent word shapes in the Wikipedia

Frequent word shapes in the Wikipedia

- (99637) x/x

Frequent word shapes in the Wikipedia

- (99637) x/x – 'predicted/known', 'comedian/writer', 'router/hub'

Frequent word shapes in the Wikipedia

- (99637) x/x – 'predicted/known', 'comedian/writer', 'router/hub'
- (91787) Xx/Xx

Frequent word shapes in the Wikipedia

- (99637) x/x – 'predicted/known', 'comedian/writer', 'router/hub'
- (91787) Xx/Xx – 'King/Prince', 'California/Beverly', 'Voutput/Vinput'

Frequent word shapes in the Wikipedia

- (99637) x/x – 'predicted/known', 'comedian/writer', 'router/hub'
- (91787) Xx/Xx – 'King/Prince', 'California/Beverly', 'Voutput/Vinput'
- (59507) x-Xx

Frequent word shapes in the Wikipedia

- (99637) x/x – 'predicted/known', 'comedian/writer', 'router/hub'
- (91787) Xx/Xx – 'King/Prince', 'California/Beverly', 'Voutput/Vinput'
- (59507) x-Xx – 'counter-Victorian', 'post-Partnership', 'al-Sarman'

Frequent word shapes in the Wikipedia

- (99637) x/x – 'predicted/known', 'comedian/writer', 'router/hub'
- (91787) Xx/Xx – 'King/Prince', 'California/Beverly', 'Voutput/Vinput'
- (59507) x-Xx – 'counter-Victorian', 'post-Partnership', 'al-Sarman'
- (54701) x-x-x

Frequent word shapes in the Wikipedia

- (99637) x/x – 'predicted/known', 'comedian/writer', 'router/hub'
- (91787) Xx/Xx – 'King/Prince', 'California/Beverly', 'Voutput/Vinput'
- (59507) x-Xx – 'counter-Victorian', 'post-Partnership', 'al-Sarman'
- (54701) x-x-x – 'process-of-discovery', 'get-rich-quick', 'million-copy-selling'

Why word shapes?

Why word shapes?

- Word shapes are important to handle unknown words.

Why word shapes?

- Word shapes are important to handle unknown words.
- Example: named entity recognition

Why word shapes?

- Word shapes are important to handle unknown words.
- Example: named entity recognition
- Recognize persons, locations, organizations etc in text

Why word shapes?

- Word shapes are important to handle unknown words.
- Example: named entity recognition
- Recognize persons, locations, organizations etc in text
- Italy's business world was rocked by the announcement last Thursday that Mr. Verdi would leave his job as vice-president of Music Masters of Milan, Inc to become operations director of Arthur Andersen.

Why word shapes?

- Word shapes are important to handle unknown words.
- Example: named entity recognition
- Recognize persons, locations, organizations etc in text
- Italy's business world was rocked by the announcement last Thursday that Mr. Verdi would leave his job as vice-president of Music Masters of Milan, Inc to become operations director of Arthur Andersen.
- LOC:Italy's business world was rocked by the announcement TIME:last Thursday that Mr. PERSON:Verdi would leave his job as vice-president of ORG:Music Masters of Milan Inc., to become operations director of ORG:Arthur Andersen.

Why word shapes?

- Word shapes are important to handle unknown words.
- Example: named entity recognition
- Recognize persons, locations, organizations etc in text
- Italy's business world was rocked by the announcement last Thursday that Mr. Verdi would leave his job as vice-president of Music Masters of Milan, Inc to become operations director of Arthur Andersen.
- LOC:Italy's business world was rocked by the announcement TIME:last Thursday that Mr. PERSON:Verdi would leave his job as vice-president of ORG:Music Masters of Milan Inc., to become operations director of ORG:Arthur Andersen.
- Word shapes can indicate that a word is a named entity and (in some cases) what type of named entity it is.

Why word shapes?

- Word shapes are important to handle unknown words.
- Example: named entity recognition
- Recognize persons, locations, organizations etc in text
- Italy's business world was rocked by the announcement last Thursday that Mr. Verdi would leave his job as vice-president of Music Masters of Milan, Inc to become operations director of Arthur Andersen.
- LOC:Italy's business world was rocked by the announcement TIME:last Thursday that Mr. PERSON:Verdi would leave his job as vice-president of ORG:Music Masters of Milan Inc., to become operations director of ORG:Arthur Andersen.
- Word shapes can indicate that a word is a named entity and (in some cases) what type of named entity it is.
- Example?

Why word shapes?

- Word shapes are important to handle unknown words.
- Example: named entity recognition
- Recognize persons, locations, organizations etc in text
- Italy's business world was rocked by the announcement last Thursday that Mr. Verdi would leave his job as vice-president of Music Masters of Milan, Inc to become operations director of Arthur Andersen.
- LOC:Italy's business world was rocked by the announcement TIME:last Thursday that Mr. PERSON:Verdi would leave his job as vice-president of ORG:Music Masters of Milan Inc., to become operations director of ORG:Arthur Andersen.
- Word shapes can indicate that a word is a named entity and (in some cases) what type of named entity it is.
- Example?
- Xx-Xx-x-Xx-x-Xx / Sainte-Euphemie-sur-Riviere-du-Sud

Why word shapes?

- Word shapes are important to handle unknown words.
- Example: named entity recognition
- Recognize persons, locations, organizations etc in text
- Italy's business world was rocked by the announcement last Thursday that Mr. Verdi would leave his job as vice-president of Music Masters of Milan, Inc to become operations director of Arthur Andersen.
- LOC:Italy's business world was rocked by the announcement TIME:last Thursday that Mr. PERSON:Verdi would leave his job as vice-president of ORG:Music Masters of Milan Inc., to become operations director of ORG:Arthur Andersen.
- Word shapes can indicate that a word is a named entity and (in some cases) what type of named entity it is.
- Example?
- Xx-Xx-x-Xx-x-Xx / Sainte-Euphemie-sur-Riviere-du-Sud
- x'Xx / o'Hara

Convert these words into word shapes

POW/MIA → Composer-in-residence →
d'Abruzzo → Aminoacylase-3 →
http://www.icssrnerc.org →

- Replace all lower case letters with 'x'
- Replace all upper case letters with 'X'
- Replace all digits letters with '9'
- "Deduplicate": any sequence of $n > 1$ identical characters c is replaced by a single copy of c

Convert these words into word shapes

POW/MIA → Composer-in-residence →
d'Abruzzo → Aminoacylase-3 →
http://www.icssrnerc.org →

- Replace all lower case letters with 'x'
- Replace all upper case letters with 'X'
- Replace all digits letters with '9'
- "Deduplicate": any sequence of $n > 1$ identical characters c is replaced by a single copy of c

Convert these words into word shapes

POW/MIA → Composer-in-residence →
d'Abruzzo → Aminoacylase-3 →
http://www.icssrnerc.org →

- Replace all lower case letters with 'x'
- Replace all upper case letters with 'X'
- Replace all digits letters with '9'
- "Deduplicate": any sequence of $n > 1$ identical characters c is replaced by a single copy of c

Convert these words into word shapes

POW/MIA → Composer-in-residence →
d'Abruzzo → Aminoacylase-3 →
http://www.icssrnerc.org →

- Replace all lower case letters with 'x'
- Replace all upper case letters with 'X'
- Replace all digits letters with '9'
- "Deduplicate": any sequence of $n > 1$ identical characters c is replaced by a single copy of c

Convert these words into word shapes

POW/MIA → X/X Composer-in-residence → Xx-x-x
d'Abruzzo → x'Xx Aminoacylase-3 → Xx-9
http://www.icssrnerc.org → x:/x.x.x

- Replace all lower case letters with 'x'
- Replace all upper case letters with 'X'
- Replace all digits letters with '9'
- "Deduplicate": any sequence of $n > 1$ identical characters c is replaced by a single copy of c

Exercise

- Provide a (theoretically possible) word shape that does not occur in Wikipedia
- The shortest such word shape wins.

- Replace all lower case letters with 'x'
- Replace all upper case letters with 'X'
- Replace all digits letters with '9'
- "Deduplicate": any sequence of $n > 1$ identical characters c is replaced by a single copy of c