Inverzní normalizace textu

Inverse Text Normalization as a Labeling Problem

Pavla Benetinová

437782

PLIN037 Sémantika a počítače

Filozofická fakulta
Masarykova univerzita

18. května 2018

Inverzní normalizace textu (INT) je standardně používaný postup pro převod mluvené řeči do psané podoby. Problematikou se zabývá také společnost Apple. Díky INT dokáže Siri číselné údaje (jako datum, čas, adresu, ...) z mluveného jazyka převést do formátovaného psaného tvaru.

mluvená forma	**psaná forma**
one forty one Dorchester Avenue Salem Massachusetts	141 Dorchester Ave., Salem, MA
set an alarm for five thirty p.m.	Set an alarm for 5:30 PM
September sixteenth twenty seventeen	September 16, 2017
twenty percent of fifteen dollars seventy three	20% of $15.73
two hundred seven point three plus six	207.3+6

Možné, leč neefektivní, způsoby řešení:

- naivní tokenizace
  - jednoduchý přepis po jedné číslovce, nepřesné
  - např. sto dvacet pět $\neq$ 100 20 5
- frázový přepis
  - často používaný u strojového překladu
  - ALE:
    - vyžaduje velké množství trénovacích dat
    - výpočetně drahé (čas, pamět)
    - nedostatek omezujících podmínek $\rightarrow$ velká chybovost
    - nevyplatí se

Siri však využívá mnohem efektivnější způsob, a to značkování, za pomoci jednoduchých gramatik. Promluva je rozdělena na jednotlivé tokeny, kterým jsou následně přiděleny značky popisující akce, které je nutné na daném tokenu provést.

Postup

1. značkování
   - přiřazení značky každému tokenu zvlášť
   - twentieth $\rightarrow$ RewriteOrdinalAsCardinalDecade_AppendComma
2. aplikace značek
   - značka specifikuje úpravy

- RewriteOrdinalAsCardinalDecade_AppendComma → 20,

3. postprocessing
   - aplikace gramatiky, pokud je třeba

Následující tabulka naznačuje body postupu 1 a 2.

mluvená forma	**značka**	**aplikace značky**
February	Default	February
twentieth	RewriteOrdinalAsCardinalDecade _AppendComma	20,
twenty	RewriteCardinalDecade	20
seventeen	RewriteCardinalTeen_SpaceNo	17

Tabulka 1: přepis, February 20, 2017

Následující tabulka naznačuje bod postupu 3.

mluvená forma	**značka**	**aplikace značky**	**psaná forma**
twenty	RewriteCardinalDecade	20	20
percent	RewritePercentSigno_SpaceN	%	%
of	Default	of	of
two	RewriteCardinal _StartMajorCurrency	<MajorCurrency>2	
hundred	RewriteMagnitudePop1 _SpaceNo	0	
five	RewriteCardinal_SpaceNo	5	
dollars	RewriteCurrencySymbol _SpaceNo_EndMajorCurrency	$<\MajorCurrency>	$205

Tabulka 2: přepis, 20% of $205

Značkování

Pro přepis velkých čísel používá Apple metodu *konečného stavového snímače*. Který zabraňuje přepisům jako: *dvě stě pět → 200 5*. A to tak, že zpětně zkontroluje, kde číslovka končí, a přiřadí

značku. V tomto případě stovkám značku *MagnitudePop1*, která přikazuje zahodit poslední 1 nulu.

Celá značka číslovky je však, viz tabulka výše, *RewriteMagnitudePop1_SpaceNo*. Kdy *Rewrite* značí, že výraz má být přepsán z textové podoby do číselné a *SpaceNo* definuje nepřidání mezery za 0, která se jinak ve výchozím nastavení přidává automaticky.

Podobná pravidla platí také pro přidávání interpunkce, lomítka, dvojtečky atd.

Postprocessing

Některé výrazy je nutné zpracovávat zpětně. Výrazy obsahující měnu. Časová vyjádření jako *25 minut do 4*. Nebo dokonce římské číslice. Pro tento způsob zpracování se využívají vytvořené gramatiky, které jsou zpracovány metodou konečného stavového snímače. V následující tabulce uvádím příklady použití gramatiky.

aplikace značek	**psaná forma**
You owe me <MajorCurrency>3$</MajorCurrency>.50	You owe me $3.50
Meet me at <RelativeTime>25 minutes to 4</RelativeTime>	Meet me at 3:35
Super Bowl <RomanNumeral>51</RomanNumeral>	Super Bowl LI

Tabulka 3: příklady použití gramatiky

Výsledky

Přesnost systému Apple ověřoval na náhodně vybraných projevů Siri, za použití postupně 500 000, 1 mil. až 5 mil. trénovacích dat. Zaměříme-li se právě na ta data, která obsahují číselné výrazy, pak přesnost za použití trénovacích dat alespoň o 1 mil. výroků přesahuje 90 %.

Závěr

V textu uvádím, na základě výchozího článku, anglické příklady funkce systému. Obdobné výsledky dostaneme ale také pro český jazyk. Ať už se jedná o časové údaje, u kterých si Siri poradí, stejně jako v angličtině, s výrazy jako například *za pět minut osm* → *7:55*. Nebo o výrazy obsahující měnu (u kterých navíc není tak složitý postup jako ohledně dolarů). Zajímavostí je, že systém nemá problém ani s údaji typu *deset kilometrů za hodinu* → *10 km/h*. Hovorový jazyk mu také nedělá potíže: *dvacet pět kilo* → *25 kg*.

Můžeme tedy tvrdit, že se jedná o velmi užitečný, technicky vyspělý a překvapivě přesný systém, který má budoucnost.

Literatura

Siri Team, 2017. Inverse Text Normalization as a Labeling Problem. *Apple Machine Learning Journal* [online]. Vol. 1, no. 3 [cit 16. 5. 2018]. Dostupné z: https://machinelearning.apple.com/2017/08/02/inverse-text-normal.html