Massively Translingual Compound Analysis and Translation Discovery

Winston Wu and David Yarowsky
Center for Language and Speech Processing, Johns Hopkins University
{wswu,yarowsky}@jhu.edu

Overview

Morphological compounding is prevalent and systematic across many of the world's languages. By examining how a particular concept is compounded across languages, can we predict how it is compounded in a specific language?

Concept		Left +	Right
ninety	zh	九 (nine)	十 (ten)
Monday	nl	maan (moon)	dag (day)
December	fi	joulu (Jule)	kuu (month)
midnight	hu	éj (night)	fél (half)
Frenchman	nl	frans (French)	man (man)
businessman	th	นัก (person)	ธุรกิ (business)
pianist	de	klavier (piano)	spieler (player)
granddaughter	no	datter (daughter)	datter (daughter)
queen	hu	király (king)	nő (woman)

By using only freely-available dictionaries, we can gather plausible compounds and construct compounding recipes. We use these recipes to discover additional compounds and identify common compounding glue characters.

Compounding Mechanisms

We consider various mechanisms to combine two components into a single compound word:

Simple concatenation: A + B
- fi: rakennustyö "construction" = construction + work
- nl: ziekenhuis "hospital" = sick + house

Glue Character: A + glue + B
- da: folk**e**afstemning "referendum" = people + vote
- nb: informasjon**s**teknologi "information technology"

Drop Left: A[:-1] + B
- de: meuchel~~n~~mörder "assassin" = assassinate + killer
- es: cant~~o~~autor "singer-songwriter" = singing + author

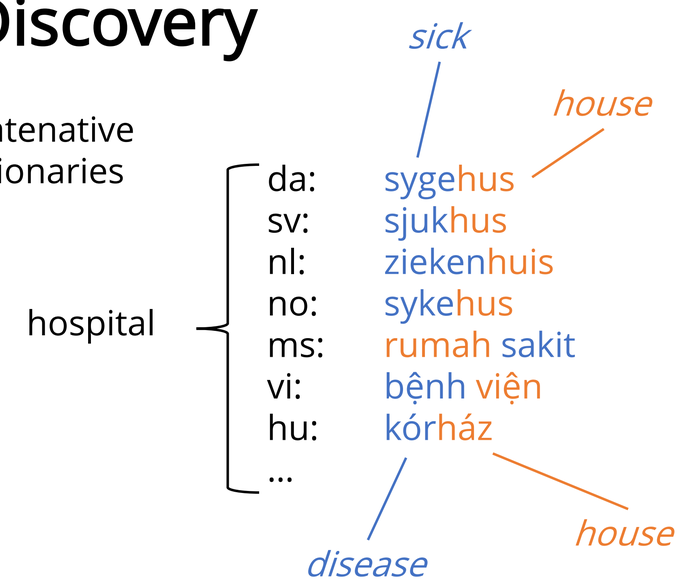
Experiments

We evaluate the effectiveness of our compound recipes in handling unknown words (held-out from the dictionary) in two directions:

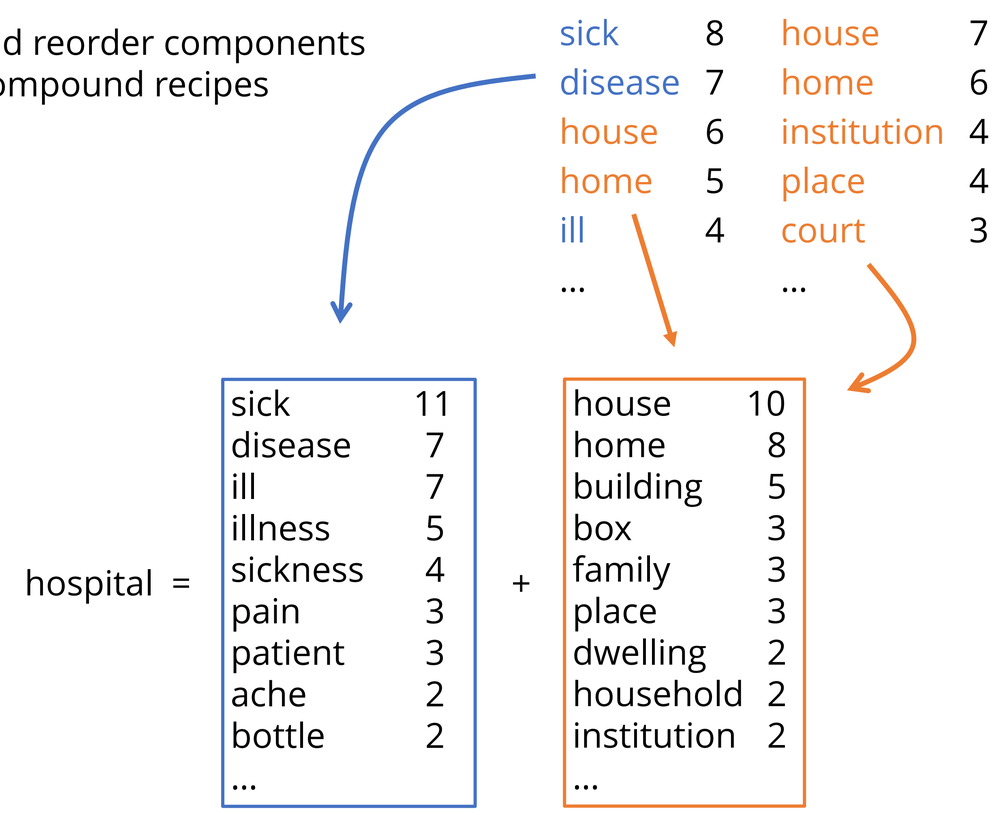
- e2f: given a concept, can we generate the foreign compound word?
- f2e: given a foreign compound word, can we analyze it and identify the concept?

Compound Discovery

1. Discover simple concatenative compounds from dictionaries



2. Cluster and reorder components to form compound recipes



3. Compound generation and augmentation

hospital → Danish

dårlig+hus *concatenation*
dårlig+hjem
dårlig+sted
...
syg+hus
syg+hjem
... *(Cartesian product)*

dårlig+e+hus *glue character*
...

dårli+hus *drop left*
...

hus+dårlig *order flip*
... *(above mechanisms)*

4. Score and search for attested terms

Compound Validity Score
$= \sqrt{\text{left\_count} \cdot \text{right\_count}}$

Score	Lng	Decomposition
12.96	nn	sjukehus = sjuk + hus ; sick + house
12.96	nl	ziekenhuis = zieken + huis ; sick + house
12.96	nl	ziekenhuis = zieke + huis ; sick + house
12.96	tpi	haus sik = haus + sik ; house + sick
12.96	no	sykehus = syk + hus ; sick + house
12.96	nb	sykehus = syk + hus ; sick + house
12.96	ms	rumah sakit = rumah + sakit ; house + sick
12.96	sv	sjukhus = sjuk + hus ; sick + house
12.96	da	sygehus = syge + hus ; sick + house
12.96	da	sygehus = syg + hus ; sick + house
12.96	id	rumah = sakit rumah + sakit ; house + sick
12.96	sv	sjukhus = sjuka + hus ; sick + house
12.96	sv	sjukhus = sjuke + hus ; sick + house
12.96	no	sykehus = syke + hus ; sick + house

Sample Results

Lang	English Concept	Foreign	Foreign Analysis	Literal Translation	Top Hyp	2nd Hyp	Rank	f2e
eo	yolk	ovoflavo	ovo + flavo	egg yellow	**yolk**	egg yolk	1	Y
wyi	skull	galk gawang	galk + ' ' + gawang	bone head	**skull**	cranium	1	Y
mns	fifty	атлов	ат + лов	five ten	fifteen	**fifty**	2	Y
nb	continent	fastland	fast + land	firm country	mainland	**continent**	2	Y

Code and data at `github.com/wswu/worcomal`