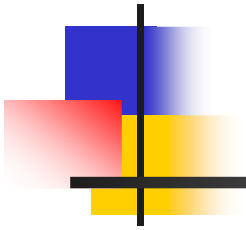


Introduction



Updated 5/2009



Today's Approach to NLP

- From ~1970-1989, people were concerned with the science of the mind and built small (toy) systems that attempted to behave intelligently.
- Recently, there has been more interest on engineering practical solutions using automatic learning (knowledge induction).
- While Chomskyans tend to concentrate on categorical judgements about very rare types of sentences, statistical NLP practitioners concentrate on common types of sentences.

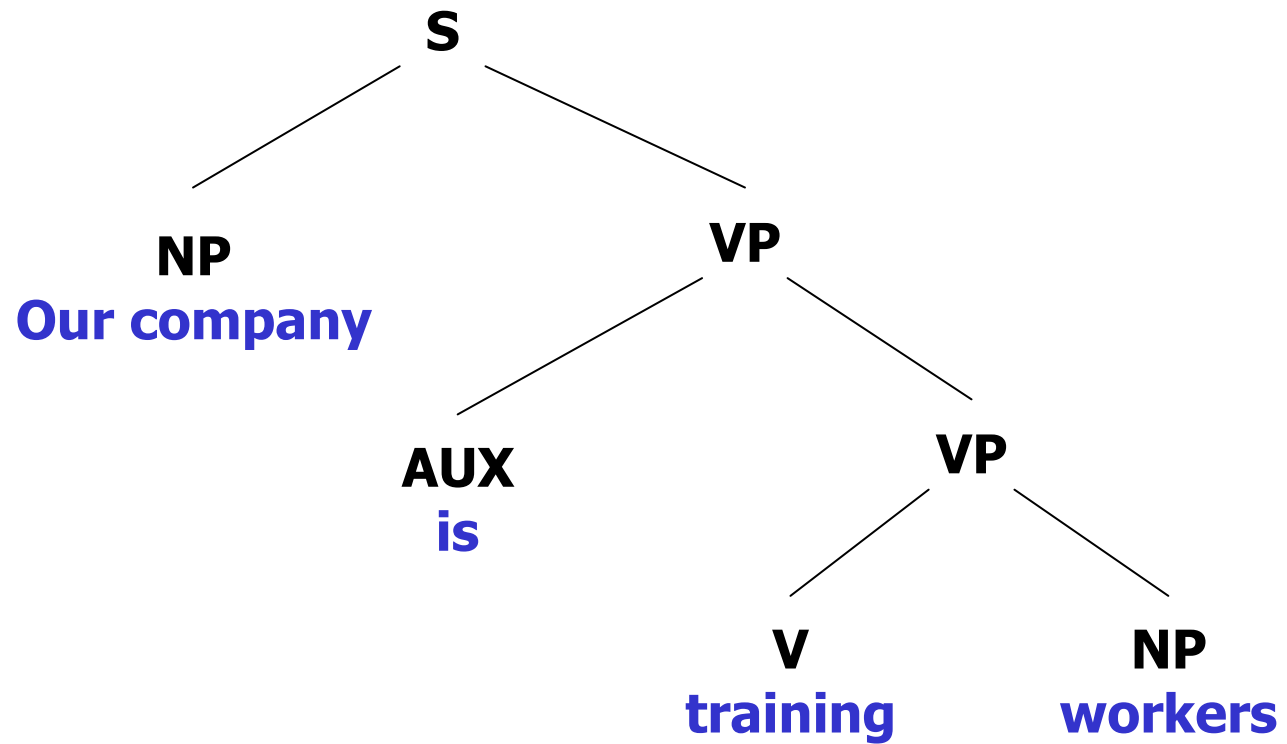


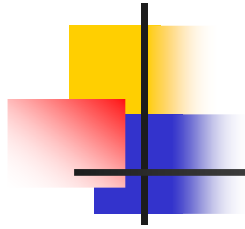
Why is NLP Difficult?

- NLP is difficult because Natural Language is highly ambiguous.
- Example: “Our company is training workers” has 3 ***parses*** (i.e., syntactic analyses).
- “List the sales of the products produced in 1973 with the products produced in 1972” has 455 parses.
- Therefore, a practical NLP system must be good at making disambiguation decisions of word sense, word category, syntactic structure, and semantic scope.

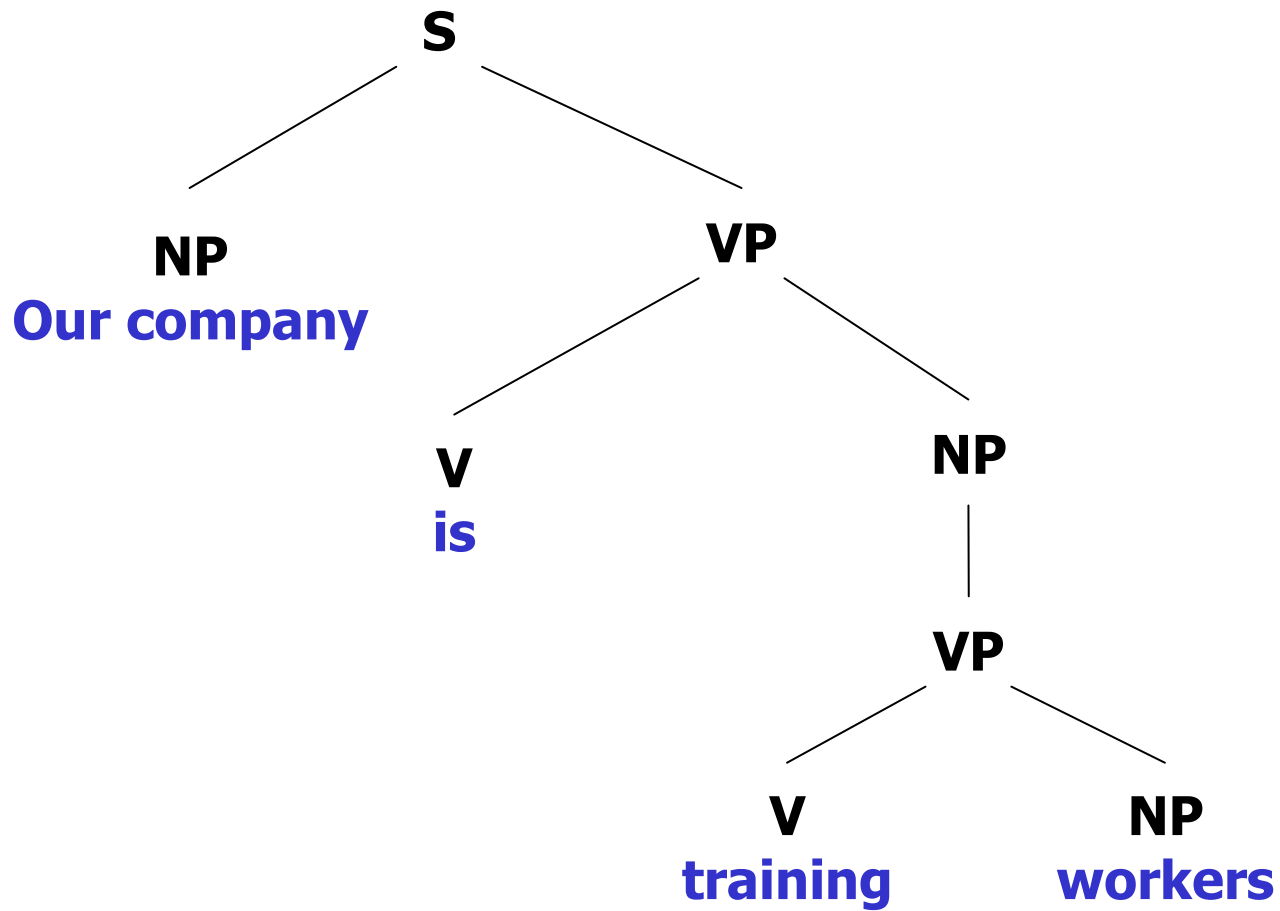


Parses I





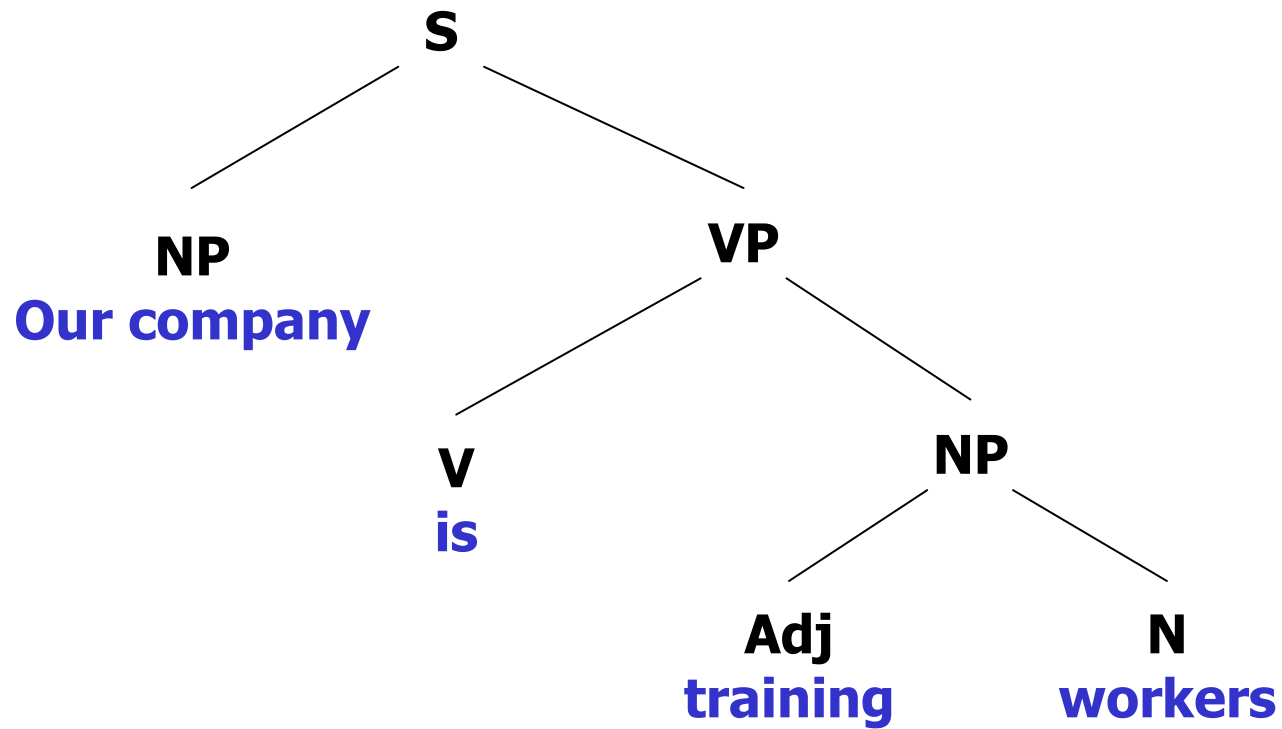
Parse II



e.g.: our problem is training workers



Parse III



e.g.: those are training wheels



Real News headlines, U.S.

- Hospitals are Sued by 7 Foot Doctors
- Astronaut Takes Blame for Gas in Spacecraft
- New Study of Obesity Looks for Larger Test Group
- Chef Throws His Heart into Helping Feed Needy
- Include your Children when Baking Cookies
- Kids Make Nutritious Snacks



Methods that don't work well

- Maximizing coverage while minimizing ambiguity is inconsistent with symbolic NLP.
- Furthermore, hand-coding syntactic constraints and preference rules are time consuming to build, do not scale up well and are brittle in the face of the extensive use of metaphor in language.
- **Example:** if we code:
- animate being --> **swallow** --> physical object
I swallowed his story, hook, line, and sinker
The supernova swallowed the planet.

What Statistical NLP can do for



us

- Disambiguation strategies that rely on hand-coding produce a knowledge acquisition bottleneck and perform poorly on naturally occurring text.
- A Statistical NLP approach seeks to solve these problems by automatically learning lexical and structural preferences from corpora. In particular, Statistical NLP recognizes that there is a lot of information in the relationships between words.
- The use of statistics offers a good solution to the ambiguity problem: statistical models are robust, generalize well, and behave gracefully in the presence of errors and new data.



Things that can be done with Text Corpora I: Word Counts

- **Word Counts to find out:**
 - What are the most common words in the text.
 - How many words are in the text (word tokens and word types).
 - What the average frequency of each word in the text is.
- **Limitation of word counts:** Most words appear very infrequently and it is hard to predict much about the behavior of words that do not occur often in a corpus. ==> Zipf's Law.

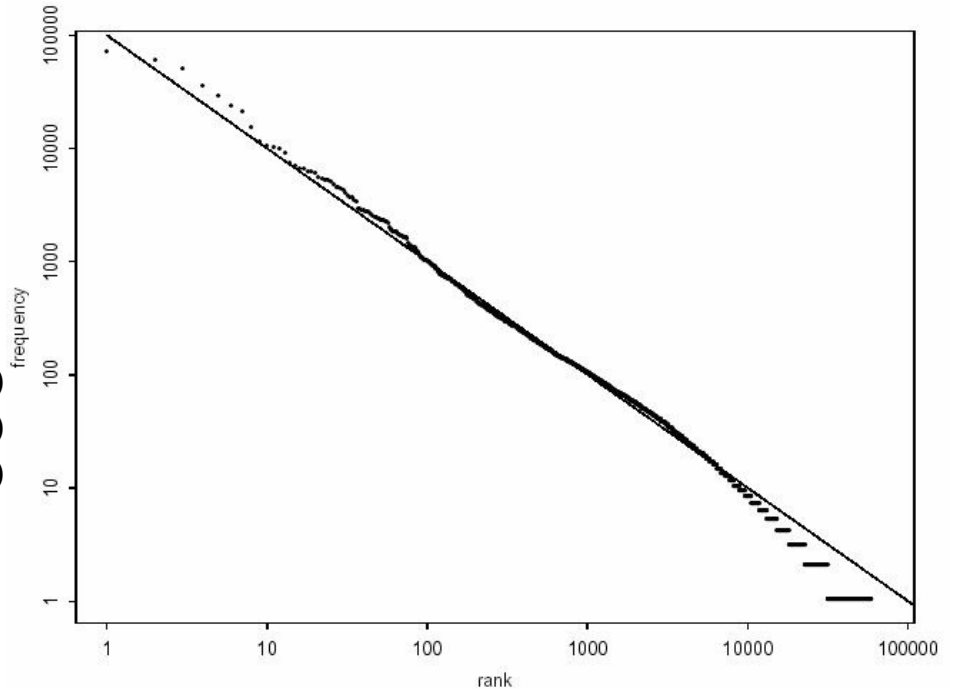


Things that can be done with Text Corpora II: Zipf's Law

- If we count up how often each word type of a language occurs in a large corpus and then list the words in order of their frequency of occurrence, we can explore the relationship between the frequency of a word, f , and its position in the list, known as its rank, r .
- Zipf's Law says that: $f \propto 1/r$
- Significance of Zipf's Law: For most words, our data about their use will be exceedingly sparse. Only for a few words will we have a lot of examples.

Zipf's law - empirical evaluation (Tom Sawyer)

Word	Freq.	Rank	f * r
the	3332	1	3332
and	2972	2	5944
a	1775	3	5235
he	877	10	8770
but	410	20	8400
be	294	30	8820
there	222	40	8880
one	172	50	8600
about	158	60	9480
more	138	70	9660
never	124	80	9920
Oh	116	90	10440
two	104	100	10400
turned	51	200	10200
you'll	30	300	9000
name	21	400	8400
comes	16	500	8000
group	13	600	7800
lead	11	700	7700
friends	10	800	8000
begin	9	900	8100
family	8	1000	8000





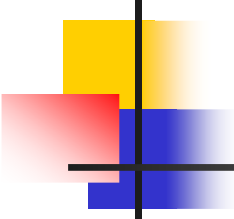
Things that can be done with Text Corpora III: Collocations

- A **collocation** is any turn of phrase or accepted usage where somehow the whole is perceived as having an existence beyond the sum of its parts (e.g., disk drive, make up, bacon and eggs).
- Collocations are important for machine translation.
- Collocation can be extracted from a text (example, the most common **bigrams** can be extracted). However, since these bigrams are often insignificant (e.g., “at the”, “of a”), they can be **filtered**.

Commonest bigram collocations in New York Times

frequency	word1	word2	frequency	word1	word2	POS	pattern
80871	of	the	11487	New	York	A N	
58841	in	the	7261	United	States	A N	
26430	to	the	5412	Los	Angeles	N N	
21842	on	the	3301	last	year	A N	
21839	for	the	3191	Saudi	Arabia	N N	
18568	and	the	2699	last	week	A N	
16121	that	the	2514	vice	president	A N	
15630	at	the	2378	Persian	Gulf	A N	
15494	to	be	2161	San	Francisco	N N	
13899	in	a	2106	President	Bush	N N	
13689	of	a	2001	Middle	East	A N	
13361	by	the	1942	Saddam	Hussein	N N	
13183	with	the	1867	Soviet	Union	A N	
12622	from	the	1850	White	House	A N	
11428	New	York	1633	United	Nations	A N	
10007	he	said	1337	York	City	N N	
			1328	oil	prices	N N	
			1210	next	year	A N	

FILTERED



Things that can be done with Text Corpora IV: Concordances

- Finding **concordances** corresponds to finding the different contexts in which a given word occurs.
- One can use a **Key Word In Context (KWIC)** concordancing program.
- Concordances are useful both for building dictionaries for learners of foreign languages and for guiding statistical parsers.



KWIC display for the word *showed* (Tom Sawyer)

1 could find a target. The librarian	“showed	off” - running hither and thither w
2 elights in. The young lady teachers	“showed	off” - bending sweetly over pupils
3 ingly. The young gentlemen teachers	“showed	off” with small scoldings and other
4 seeming vexation). The little girls	“showed	off” in various ways, and the littl
5 n various ways, and the little boys	“showed	off” with such diligence that the a
6 t genuwyne?” Tom lifted his lip and	showed	the vacancy. “Well, all right,” sai
7 is little finger for a pen. Then he	showed	Huckleberry how to make an H
8 ow's face was haggard, and his eyes	showed	the fear that was upon him. Whe
9 not overlook the fact that Tom even	showed	a marked aversion to these inquest
10 wn. Two or three glimmering lights	showed	where it lay, peacefully sleeping,
11 ird flash turned night into day and	showed	every little grass-blade, separate
12 that grew about their feet. And it	showed	three white, startled faces, too. A
13 he first thing his aunt said to him	showed	him that he had brought his sorro
14 p from her lethargy of distress and	showed	good interest in the proceedings. S
15 ent a new burst of grief from Becky	showed	Tom that the thing in his mind ha



‘Showed’ syntactic structures

- NP_{agent} *showed* NP_{receptient}
- NP_{agent} *showed off* PP