Chapter 1 Introduction

**Phonology**: how words are pronounced in terms of sequences of sounds & how each of these sounds is realized acoustically.

**Morphology**: the way words break down into component parts that carry meanings (e.g. singular vs. phural).

**Lexical semantics**: the meaning of all the words.

**Compositional semantics**: the meaning of words when combining with other words.

Request & Statement & Information Question

Coreference Resolution: the meaning of pronouns referencing to previous contents.

Ambiguity: Multiple linguistic structures can be built for the input.

**Part-of-speech Tagging**: deciding whether “duck” is a verb or noun.

**Word Sense Disambiguation**: deciding whether “make” means “create” or “cook”.

**Syntactic Disambiguation**: deciding whether “her” and “duck” belong to the same entity (can be addressed by probabilistic parsing).

**Speech Act Interpretation**: deciding whether a sentence is statement or question.

(all are important kinds of Lexical Disambiguation)

\*State Space Search algorithms (e.g. Dynamic Programming)

\*Machine Learning algorithms: (e.g. Expectation-Maximization)

Models:

State Machines (e.g. FSM) & Grammars (e.g. Regex) & Logic (e.g. FOL) & Probabilistic (e.g. HMM) & Vector-Space (e.g. Linear Algebra)

Departments:

Computational linguistics, Natural language processing, Speech recognition, Computational psycholinguistics

Paradigms:

Stochastic (e.g. HMM) & Logic-based (e.g. Q-system) \* Natural Language Understanding (e.g. SHRDLU) & Discourse Modeling (Belief-Desire-Intention)

Chapter 4 N-grams

1. grams can be applied to:

Part-of-speech Tagging, Natural Language Generation, Word Similarity, Authorship Identification, Sentiment Extraction, Predictive Text Input

Corpus: online collection of text or speech.

“I do uh main- mainly business data processing”

**Fragment**: the broken-off word “main-”

**Fillers/Filled Pauses**: words like “uh”, “um”

(both are disfluencies)

“cats” vs. “cat”: same **lemma (cat)**, different **wordforms (singular/plural)**

Types: the number of distinct words in a corpus or vocabulary size V

Tokens: the total number of running words

Open/Closed vocabulary (existance of unknown words)

**Intrinsitic evaluation** measures the quality of a model independent of any application. **Perplexity** is the most common one for N-gram.

Perplexity (PP) = P(w1w2...wN)-1/N

(weighted average branching factor)

Smoothing: modification that address the poor estimates that are due to variability in small datasets. Laplace smoothing is a simple one.

PLaplace(wi) = (ci + 1) / (N + V)

**Discounting**: smoothing by lowering some non-zero counts in order to get the probability mass that will be assigned to the zero counts.

Chapter 5 Word Classes and POS Tagging

Parts of Speech: noun（名词）, verb（动词）, pronoun（代词）, preposition（介词）, adverb（副词）, conjunction（连词）, participle（小品词-与动词构成短语动词的副词或介词）, article （定冠词）

- Closed class: parts that have relatively fixed membership (e.g. article in English)

- Open class: parts that are continually coined or borrowed from other languages

**Purposes** of POS Tagging:

inference of the pronounciation of words, stemming for informational retrieval

**Methods** of POS Tagging:

rule-based tagging, HMM tagging, maximum entropy tagging, transformation-based tagging, memory-based tagging

Proper nouns: names of specific persons or entities

Common nouns: general names usually preceded by articles

POS Tagging is usually performed after tokenization.

1. Rule-based POS Tagging

E.g. ADVERBIAL-THAT RULE

Given input: “that”

If (+1 A/ADV/QUANT); (+2 SENT-LIM); (NOT -1 SVOC/A);

Then eliminate non-ADV tags

Else eliminate ADV tag

1. HMM POS Tagging

Choose the tag sequence which is most probable given the observation sequence of n words.

t = argmax P(t|w) = argmax P(w|t)P(t)/P(w) ~ argmax P(w|t)P(t)

HMM with Deleted-Interpolation

1. Transformation-based Tagging

Label every word with its most-likely tag.

Chapter 9 Automatic Speech Recognition

Automatic Speech Recognition: mapping from an acoustic signal to a string of words.

Automatic Speech Understanding: produce understanding of the sentence.

Large-Vocabulary Continuous Speech Recognition: system that have 20000-60000 words vocabulary, with words running together naturally (continuous), speaker independent.

(The rest focuses on acoustic and thus skipped)