Морфологическая обработка в задачах информационного поиска

Павел Браславский

Зачем нужна морфология в поиске?

- Классы эквивалентности ключевых слов при поиске:

кошка, кошки, кошку, кошкой, кошке…

- Извлечение информации (information extraction)
- Учет при ранжировании
- «примитивный синтаксис»

Типы анализа

- Стемминг – выделение основы

лесной, лес, лесистый, леса → лес

система, системный, систематизировать → систем

- Приведение к словарной форме

лесного, лесному → лесной

леса → лес

танцующая → танцевать

Типы анализа – 2

- POS-tagging (part-of-speech)

Танцующая <V> в <PREP> темноте <N>

- Полная морфологическая информация

Танцующая <V, прич, несоверш, наст., ед., жен., им.> в <PREP> темноте <N, жен., неод., ед., предл.>

Части речи	Грамматические категории (в скобках приведены сокращенные названия их значений)
Существительное (сущ.)	Род (м., ж., ср., $\overline{р}$), число (ед., мн.), падеж (им., род., дат., вин., тв., пр.), одушевленность (од., неод., $\overline{о}$)
Полное прилагательное (прил. полн.)	Пассивность (пасс., акт., $\overline{п}$), время (прош., наст., $\overline{в}$), род (м., ж., ср., $\overline{р}$), число (ед., мн.), падеж (им., род., дат., вин., тв., пр.), одушевленность (од., неод., $\overline{о}$), вид (сов, нсв, $\overline{вид}$)
Краткое прилагательное (прил. кр.)	Пассивность (пасс., акт.), время (прош., наст., буд.,), род (м., ж., ср., $\overline{р}$), число (ед., мн.), вид (сов., нсв., $\overline{вид}$)
Глагол (глаг.)	Пассивность (акт., пасс.), время (прош., наст., буд.), род. (м., ж., ср., $\overline{р}$), число (ед., мн.), вид (сов, нсв, $\overline{вид}$)
Инфинитив (инф.)	Пассивность (акт., пасс.), род (м., ж., ср., $\overline{р}$), число (ед., мн., $\overline{ч}$)
Деепричастие (деепр.)	Пассивность (акт., пасс.), время (прош., наст.), вид (сов, нсв, $\overline{вид}$)
Наречие (нареч.)	Тип: обстоятельственное (обст.), определительное (опр.)
Количественное числительное (числ.)	Тип: «1», «2», «5», дробное (дробн.), неопределенное (неопр.), именованное (именов.)
Местоимение (мест.)	Класс: притяжательное (прит.), указательное (указ.), возвратное (возвр.), возвратно-атрибутивное (возвр.-атр.), третьего лица (3 л.); падеж (им., род., дат., вин., тв., пр., $\overline{п}$), число (ед., мн., $\overline{ч}$), род (м., ж., ср., $\overline{р}$)
Союз	Тип: сочинительный (соч.), подчинительный (подч.)
Предлог (предл.)	Падеж (род., дат., вин., тв., пр.)
Частица	Тип: вопросительная (вопр.), отрицательная (отр.)
Синтаксический знак (синт. зн.)	—

Попов, 1982

Грамматическая омонимия

Объект анализа – отдельное слово → неоднозначность

падали → падаль (сущ.), падать (гл.)

печь → печь (сущ.), печь (гл.)

черепах → череп (сущ., муж. род), черепаха (сущ., жен. род.)

стекла → стекло (сущ.), стекать (гл.)

ученый → учить (гл.), ученый (сущ.)

Английский: 1,2 – 1,5 тэга на словоформу

Пример: mystem

Он сделал это так неловко, что задел образок моего ангела, висевший на дубовой спинке кровати, и что убитая муха упала мне прямо на голову.

Л.Н.Толстой, «Детство»

Он{он=S,сред,неод=(им,ед|им,мн|род,ед|род,мн|дат,ед|дат,мн|вин,ед|вин,мн|твор,ед|твор,мн|пр,ед|пр,мн)|он=S,ед,муж,од=им}

сделал{сделать=V,сов=прош,ед,изъяв,муж}

это{это=S,ед,сред,неод=(им|вин)|этот=A=(им,ед,сред|вин,ед,сред)|это=PART=}

так{так=ADV=|так=PART=|так=CONJ=}

неловко{неловкий=A=ед,кр,сред|неловко=ADV=}

что{что=CONJ=|что=S,ед,сред,неод=(им|вин)}

задел{задевать=V=прош,ед,изъяв,муж,сов|задел=S,муж,неод=(им,ед|вин,ед)}

образок{образок=S,муж,неод=(им,ед|вин,ед)}

моего{мой=A=(род,ед,муж|род,ед,сред|вин,ед,муж,од)}

ангела{ангел=S,муж,од=(род,ед|вин,ед)}

висевший{висеть=V,несов=(прош,им,ед,прич,муж|прош,вин,ед,прич,муж,неод)}

на{на=PR=|на=PART=}

дубовой{дубовый=A=(род,ед,жен|дат,ед,жен|твор,ед,жен|пр,ед,жен)|дубова=S,жен,од=(род,ед|дат,ед|твор,ед|пр,ед)}

спинке{спинка=S,жен,неод=(дат,ед|пр,ед)}

Методы

- Процедурный
- Табличный
- Статистический
- Различные комбинации

Алгоритм Портера

- Самый распространенный стеммер для английского языка
- 5 циклов усечения
- Каждый цикл – набор команд
- В первую очередь выполняется операция над самым длинным суффиксом

Алгоритм Портера – фрагмент

- *sses* $\rightarrow$ *ss*
- *ies* $\rightarrow$ *i*
- *ational* $\rightarrow$ *ate*
- *tional* $\rightarrow$ *tion*

- Weight of word sensitive rules
- *(m>1) EMENT* →
  - *replacement* → *replac*
  - *cement* → *cement*

Manning&Raghavan, 2005

Попов, 1982

Морфологические типы существительных

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Им. ед.	∅	∅	∅	а	я, ья	я	я	ь	ь, й	й	о	о	я	е	е, ье	е
Род. ед.	а	а	а	ы, и	и, ьи	и	и	и	я	я	а	а	и	а	я, ья	я
Дат. ед.	у	у	у	е	е, ье	е	и	и	ю	ю	у	у	и	у	ю, ью	ю
Вин. ед.				у	ю, ью	ю	ю	ь								
Тв. ед.	ом, ем	ом	ом	ой, ой	ей, ьей	ей	ей	ью	ем	ем	ом	ем	ем	ем	ем, ьем	ем
Предл. ед.	е	е	е	е	е, ье	е	и	и	е	и	е	е	и	е	е, ье	и
Им. мн.	ы, и, ья	ы, и	а	ы, и	и, ьи	и	и	и	и	и	а	и, ья	а	а	я, ье	я
Род. мн.	ов, ев, ей, ьев	∅	ов	∅	ей	й, ь, ∅	й	ей	ей, ев	ев	∅, ов	∅, ов, ьев	∅	∅, ев	ий, ей	й
Дат. мн.	ам, ьям	ам	ам	ам	ям, ьям	ям	ям	ям	ям	ям	ам	ам, ьям	ам	ам	ям, ьям	ям
Вин. мн.																
Тв. мн.	ами, ьями	ами	ами	ами	ьями, ями	ями	ями	ями	ями	ями	ами	ьями, ами	ами	ами	ями, ьями	ями
Предл. мн.	ах. ьях	ах	ах	ах	ьях, ях	ях	ях	ях	ях	ях	ах	ах, ьях	ах	ах	ях, ьях	ях
Примеры	Процесс, брат	Грамм	Номер	Матрица	Ступня, статья	Идея	Линия	Мишень	Забой, уголь	Санаторий	Тело	Дерево	Время	Полотенце	Поле, устье	Влияние

Процедурный подход

- Словарь основ
- Словарь готовых форм (СГФ)

- Поиск в СГФ
- Выделение основы
- Поиск основы в словаре

(см. Попов, 1982, с. 234-235)

Табличный подход

- волка → волк (муж., од.; ед.ч., [р.п.|в.п.])
- не → не (частица)
- корми → кормить (несоверш.; повел. накл., ед. ч.)
- в → в (предлог)

Как сформировать таблицу?

Зализняк А.А. Грамматический словарь русского языка

~100 тыс. входов

Модель русского словоизменения

Пример: лев мо 1*b (животное)

лев м 1a (денежная единица)

стричь нсв 8b (-г-)

прихожая ж (п 4a)

Основа большинства машинных морфологий РЯ

Автомобилестроения (мн.ч.), деревяннее (ср.), при → пря

Как быть с новыми словами?

- *Варкалось, хливкие шорьки…*
- *Мерчерндайзер, бурбулятор*

Статистический стеммер

словарями → словарь → словар-ями → ар-ями

топорами → топор → топор-ами → ор-ами

летящего → лететь → лет-ящего → ет-ящего

летящего → летящий → летящ-его → ящ-его

+ правило: одна гласная в основе

имя	ра	546
има	ро	154
огещя	те	12
оге	щя	12