

פתרון רב-משמעויות של Shallow Parsing על-ידי ניתוח תכונות קוליות

מוטיבציה:

בדיבור יש מידע לא ורבלי (אינטונציה, הפסקות, ...) שאינו קיים בטקסט כתוב. במסגרת הפרויקט ננסה לנצל מידע זה להבנת הדיבור. יעמדו לרשותכם תוכנות לעיבוד קול. תשתמשו בשיטות לימוד מכונה (Machine Learning) ובינה מלאכותית. התוכנית שתכתבו תלמד מדוגמאות מתויגות ואתם תיצרו את הדוגמאות. הפרויקט עוסק באחד ההיבטים בתחום "הבנת דיבור". שיטת ה-Chunking היא אחת השיטות ל-Shallow Parsing והיא יכולה לשמש כבסיס להבנת דיבור. בשיטה זו לא מנסים לקבל ניתוח מלא של משפט, אלא מנסים לזהות יחידות משמעות יותר בסיסיות כמו Verb Phrase, Noun Phrase, וכיו. את המשפט

[[Old men] and boys] [wear [blue jeans]]

ניתן לחלק ל-chunks לפי החלוקה בסוגריים:

- Old men
- Boys
- Old men and boys
- blue jeans
- wear blue jeans

מטרת הפרויקט היא לזהות כמה שאפשר יותר מבני סוגריים על סמך תכונות פרוזודיות, כלומר, לבחור את הניתוח הנכון, למשל, לא לבחור Old [men and boys]

תיאור הפרויקט:

בפרויקט תידרשו לכתוב תוכנה עם:

- קלט אימון:
 - אוסף קבצי דיבור וטקסט מתויגים למטרת לימוד -
 - כל קובץ דיבור יכיל משפט מדובר;
 - לכל קובץ דיבור יהיה מספר קבצי טקסט: כל אחד מהקבצים מכיל צורת chunking שונה של קובץ הדיבור. כל chunk יצויין על ידי זמן התחלתו וסיומו.
 - לכל קובץ דיבור יהיה תג שיצייין את צורת ה-chunking הנכונה.
- קלט לשלב הבדיקה:
 - קובץ דיבור שמכיל משפט מדובר;
 - מספר קבצי טקסט: כל אחד מהקבצים מכיל צורת chunking שונה של קובץ הדיבור.
- פלט לשלב הבדיקה:
 - קובץ שמכיל תג שיצייין אחת צורות ה-chunking מבין הצורות הנתונות שהיא הצורה הנכונה לדעת התוכנית;
 - רמת הביטחון: מספר בין 0 ל-1 המצייין רמת הביטחון של התוכנה בהחלטתה.

אמצעי המחקר:

- PRAAT: תוכנה לעיבוד קול
- C4.5: תוכנה עבור Machine Learning
- JAVA: את התוכנה תכתבו ב-JAVA
- ASR: Microsoft Speech SDK 5.1 מתוך Microsoft Speech Recognizer

ספרות:

ספר:
כותרת:

"Spoken Language Processing,
A Guide to Theory, Algorithm, and System Development"

מחברים: Xuedong Huang, Alex Acero, Hsiao-Wuen Hon
פרק 17, עמודים 853-866