

پیکره و برجسب‌گذاری آن با کمک رایانه: ویژگی‌ها، کاستی‌ها و چالش‌ها

مسعود قیومی
M.Ghayoomi@ihcs.ac.ir

پژوهشکده زبان‌شناسی
پژوهشگاه علوم انسانی و مطالعات فرهنگی

۲۵ دی ۱۳۹۸

فهرست موضوعات

۱ مقدمه

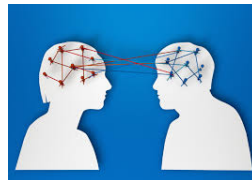
۲ پیکره زبانی و زبان‌شناسی پیکره‌ای

۳ نشانه‌گذاری داده

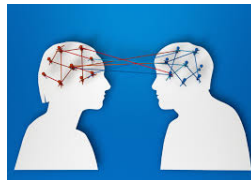
۴ چالش‌ها

۵ جمع‌بندی

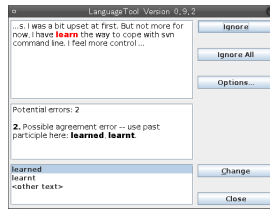
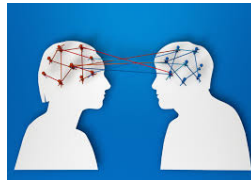
زبان‌شناسی رایانشی



زبان‌شناسی رایانشی



زبان‌شناسی رایانشی



فهرست موضوعات

مقدمه

۱

پیکره زبانی و زبان‌شناسی پیکره‌ای

۲

نشانه‌گذاری داده

۳

چالش‌ها

۴

جمع‌بندی

۵

پیکره زبانی (I)

مجموعه‌ای از داده متنی یا صوتی الکترونیکی شده به عنوان یک منبع زبانی برای تحلیل‌های زبان‌شناختی

پیکره زبانی (I)

- مجموعه‌ای از داده متنی یا صوتی الکترونیکی شده به عنوان یک منبع زبانی برای تحلیل‌های زبان‌شناختی
- «زبان‌شناسی پیکره‌ای» به اصول و شیوه‌های به کارگیری پیکره در مطالعات زبان‌شناختی می‌پردازد (کریستال، ۱۹۹۲)

پیکره زبانی (I)

- مجموعه‌ای از داده متنی یا صوتی الکترونیکی شده به عنوان یک منبع زبانی برای تحلیل‌های زبان‌شناختی
- «زبان‌شناسی پیکره‌ای» به اصول و شیوه‌های به کارگیری پیکره در مطالعات زبان‌شناختی می‌پردازد (کریستال، ۱۹۹۲)
- ویژگی‌های پیکره (مکانری و ویلسون، ۱۹۹۶)
- نماینده‌ی زبان: حاوی نمونه‌های متنوع زبانی بدون هرگونه یکسوشدگی

پیکره زبانی (I)

- مجموعه‌ای از داده متنی یا صوتی الکترونیکی شده به عنوان یک منبع زبانی برای تحلیل‌های زبان‌شناختی
- «زبان‌شناسی پیکره‌ای» به اصول و شیوه‌های به کارگیری پیکره در مطالعات زبان‌شناختی می‌پردازد (کریستال، ۱۹۹۲)
- ویژگی‌های پیکره (مکانری و ویلسون، ۱۹۹۶)
- نماینده‌ی زبان: حاوی نمونه‌های متنوع زبانی بدون هرگونه یکسودگی
- متوازن بودن داده‌های انتخاب‌شده برای پوشش انواع مختلف متن

پیکره زبانی (I)

- مجموعه‌ای از داده متنی یا صوتی الکترونیکی شده به عنوان یک منبع زبانی برای تحلیل‌های زبان‌شناختی
- «زبان‌شناسی پیکره‌ای» به اصول و شیوه‌های به کارگیری پیکره در مطالعات زبان‌شناختی می‌پردازد (کریستال، ۱۹۹۲)
- ویژگی‌های پیکره (مکانری و ویلسون، ۱۹۹۶)
- نماینده‌ی زبان: حاوی نمونه‌های متنوع زبانی بدون هرگونه یکسودگی
- متوازن بودن داده‌های انتخاب‌شده برای پوشش انواع مختلف متن
- معیارهای انتخاب متن از نظر حالت متن (گفتاری یا شنیداری) یا نوع متن (کتاب، مجله، روزنامه و غیره)

پیکره زبانی (I)

- مجموعه‌ای از داده متنی یا صوتی الکترونیکی شده به عنوان یک منبع زبانی برای تحلیل‌های زبان‌شناختی
- «زبان‌شناسی پیکره‌ای» به اصول و شیوه‌های به کارگیری پیکره در مطالعات زبان‌شناختی می‌پردازد (کریستال، ۱۹۹۲)
- ویژگی‌های پیکره (مکانری و ویلسون، ۱۹۹۶)
- نماینده‌ی زبان: حاوی نمونه‌های متنوع زبانی بدون هرگونه یکسودگی
- متوازن بودن داده‌های انتخاب‌شده برای پوشش انواع مختلف متن
- معیارهای انتخاب متن از نظر حالت متن (گفتاری یا شنیداری) یا نوع متن (کتاب، مجله، روزنامه و غیره)
- محدودیت در اندازه نمونه‌های جمع‌آوری شده

پیکره زبانی (II)

- ◀ سایر ویژگی‌ها
- ◀ بازه زمانی: معاصر یا تاریخی

پیکره زبانی (II)

◀ سایر ویژگی‌ها

◀ بازه زمانی: معاصر یا تاریخی

◀ سبک داده: رسمی/ غیررسمی؛ فرامعیار/ معیار / زیرمعیار (داگلاس، ۲۰۰۳)

پیکره زبانی (II)

- ◀ سایر ویژگی‌ها
- ◀ بازه زمانی: معاصر یا تاریخی
- ◀ سبک داده: رسمی/ غیررسمی؛ فرامعیار/ معیار / زیرمعیار (داگلاس، ۲۰۰۳)
- ◀ اندازه پیکره: متشکل از تمام یا بخشی از متن

پیکره زبانی (II)

◀ سایر ویژگی‌ها

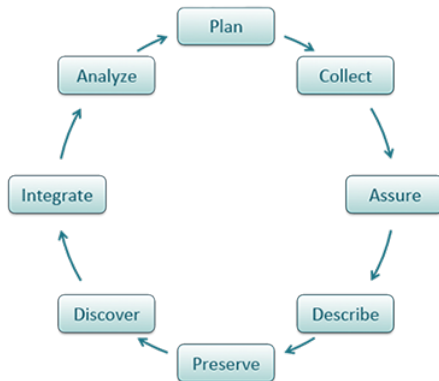
- ◀ بازه زمانی: معاصر یا تاریخی
- ◀ سبک داده: رسمی/ غیررسمی؛ فرامعیار/ معیار / زیرمعیار (داگلاس، ۲۰۰۳)
- ◀ اندازه پیکره: متشکل از تمام یا بخشی از متن
- ◀ دامنه پیکره: پیکره باز (پیکره عمومی)، پیکره خاص حاوی یک محتوای مشخص در حوزه پزشکی، علوم رایانه، شیمی، زبان‌آموزی و غیره

پیکره زبانی (II)

◀ سایر ویژگی‌ها

- ◀ بازه زمانی: معاصر یا تاریخی
- ◀ سبک داده: رسمی/ غیررسمی؛ فرامعیار/ معیار / زیرمعیار (داگلاس، ۲۰۰۳)
- ◀ اندازه پیکره: متشکل از تمام یا بخشی از متن
- ◀ دامنه پیکره: پیکره باز (پیکره عمومی)، پیکره خاص حاوی یک محتوای مشخص در حوزه پزشکی، علوم رایانه، شیمی، زبان‌آموزی و غیره
- ◀ زبان پیکره: یک زبان، دو یا چند زبان (موازی/تطبیقی)

چرخه داده



<https://www.dataone.org/data-management-planning>

بایگانی داده

◀ استفاده از منابع الکترونیکی آماده در یک سازمان



جمع‌سپاری

◀ مشارکت افراد جامعه در انجام یک کار یا پیشبرد یک هدف



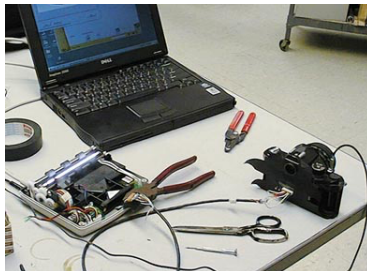
داده به اشتراک گذاشته شده

◀ استفاده از منابع به اشتراک گذاشته شده از طریق وب



داده خاص

◀ تهیه هدفمند داده باتوجه‌به نیاز



فهرست موضوعات

مقدمه

۱

پیکره زبانی و زبان‌شناسی پیکره‌ای

۲

نشانه‌گذاری داده

۳

چالش‌ها

۴

جمع‌بندی

۵

نشانه‌گذاری دستی

- استفاده از انسان خبره برای نشانه‌گذاری داده زمانی که هیچ‌گونه داده برچسب‌گذاری شده با اهداف مورد نیاز موجود نیست

نشانه‌گذاری دستی

- استفاده از انسان خبره برای نشانه‌گذاری داده زمانی که هیچ‌گونه داده برچسب‌گذاری شده با اهداف مورد نیاز موجود نیست
- نقطه قوت
- خلق چیزی که قبلاً وجود نداشته‌است



نشانه‌گذاری دستی

◀ استفاده از انسان خبره برای نشانه‌گذاری داده زمانی که هیچ‌گونه داده
برچسب‌گذاری شده با اهداف مورد نیاز موجود نیست

◀ نقطه قوت

◀ خلق چیزی که قبلاً وجود نداشته‌است

◀ کیفیت بالا در صحت برچسب‌ها به دلیل تهیه توسط
ناظر انسانی



نشانه‌گذاری دستی

- استفاده از انسان خبره برای نشانه‌گذاری داده زمانی که هیچ‌گونه داده برچسب‌گذاری شده با اهداف مورد نیاز موجود نیست



- نقطه قوت

خلق چیزی که قبلاً وجود نداشته‌است

کیفیت بالا در صحت برچسب‌ها به دلیل تهیه توسط ناظر انسانی



- نقطه ضعف

زمان بر و طاقت‌فرسا بودن



نشانه‌گذاری دستی

◀ استفاده از انسان خبره برای نشانه‌گذاری داده زمانی که هیچ‌گونه داده
 برچسب‌گذاری شده با اهداف مورد نیاز موجود نیست

◀ نقطه قوت

◀ خلق چیزی که قبلاً وجود نداشته‌است

◀ کیفیت بالا در صحت برچسب‌ها به دلیل تهیه توسط
 ناظر انسانی



◀ نقطه ضعف

◀ زمان‌بر و طاقت‌فرسا بودن

◀ پرهزینه بودن



نشانه‌گذاری دستی

- استفاده از انسان خبره برای نشانه‌گذاری داده زمانی که هیچ‌گونه داده برچسب‌گذاری شده با اهداف مورد نیاز موجود نیست



- نقطه قوت

- خلق چیزی که قبلاً وجود نداشته‌است
- کیفیت بالا در صحت برچسب‌ها به دلیل تهیه توسط ناظر انسانی



- نقطه ضعف

- زمان بر و طاقت‌فرسا بودن
- پرهزینه بودن
- تهیه حجم محدود داده



نشانه‌گذاری دستی

- استفاده از انسان خبره برای نشانه‌گذاری داده زمانی که هیچ‌گونه داده برچسب‌گذاری شده با اهداف مورد نیاز موجود نیست



- نقطه قوت

- خلق چیزی که قبلاً وجود نداشته‌است
- کیفیت بالا در صحت برچسب‌ها به دلیل تهیه توسط ناظر انسانی

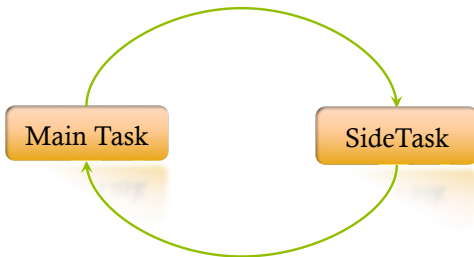


- نقطه ضعف



- زمان بر و طاقت‌فرسا بودن
- پرهزینه بودن
- تهیه حجم محدود داده
- ممکن است تنوع در نشانه‌گذاری داده‌ها اتفاق بیفتد و به عدم انسجام در داده و کاهش کیفیت نشانه‌گذاری بیانجامد.

استفاده از روش خودایستا (I)



استفاده از روش خودایستا (II)

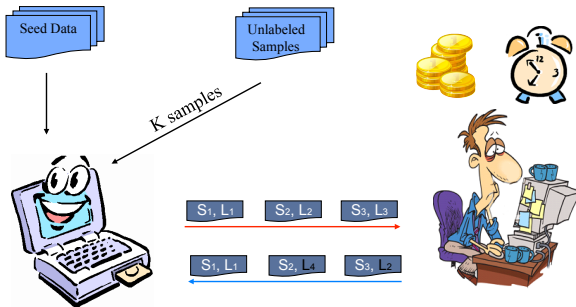
Muench-hausen Baron



استفاده از روش خودایستا (III)



استفاده از روش خودایستا (IV)



استفاده از روش خودایستا (V)

- ◀ نقطه قوت
- ◀ کاربرد ماشین برای نشانه‌گذاری داده

استفاده از روش خودایستا (V)

◀ نقطه قوت

◀ کاربرد ماشین برای نشانه‌گذاری داده

◀ کاهش نه‌چندان زیاد تلاش انسان برای نشانه‌گذاری داده

استفاده از روش خودایستا (V)

◀ نقطه قوت

- ◀ کاربرد ماشین برای نشانه‌گذاری داده
- ◀ کاهش نه‌چندان زیاد تلاش انسان برای نشانه‌گذاری داده
- ◀ کاهش نه‌چندان زیاد در صرف زمان و هزینه برچسب‌گذاری

استفاده از روش خودایستا (V)

◀ نقطه قوت

- ◀ کاربرد ماشین برای نشانه‌گذاری داده
- ◀ کاهش نه‌چندان زیاد تلاش انسان برای نشانه‌گذاری داده
- ◀ کاهش نه‌چندان زیاد در صرف زمان و هزینه برچسب‌گذاری

◀ نقطه ضعف

- ◀ نیاز به دخالت انسان در نشانه‌گذاری داده

استفاده از روش خودایستا (V)

◀ نقطه قوت

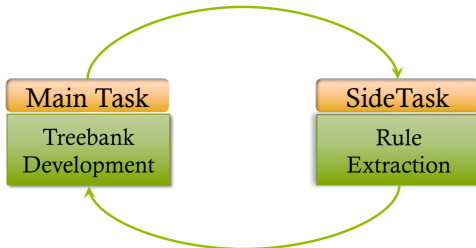
- ◀ کاربرد ماشین برای نشانه‌گذاری داده
- ◀ کاهش نه‌چندان زیاد تلاش انسان برای نشانه‌گذاری داده
- ◀ کاهش نه‌چندان زیاد در صرف زمان و هزینه برچسب‌گذاری

◀ نقطه ضعف

- ◀ نیاز به دخالت انسان در نشانه‌گذاری داده
- ◀ نیاز به نشانه‌گذاری تمام داده

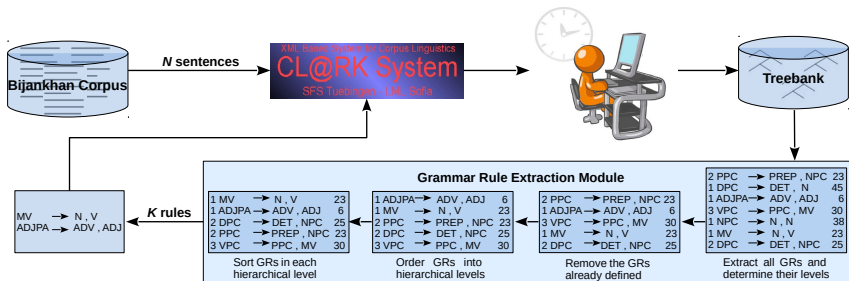
استفاده از روش خودایستا (VI)

◀ قیومی (۲۰۱۲b)



استفاده از روش خودایستا (VII)

❖ قیومی (۲۰۱۲ا)



استفاده از یادگیری فعال

◀ سِتِلِس (۲۰۱۲)

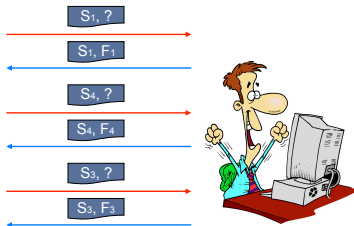
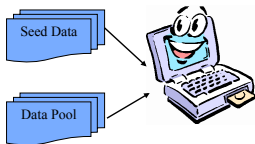
◀ نوعی یادگیری ماشینی کم‌نظارت

استفاده از یادگیری فعال

◀ سِئِلس (۲۰۱۲)

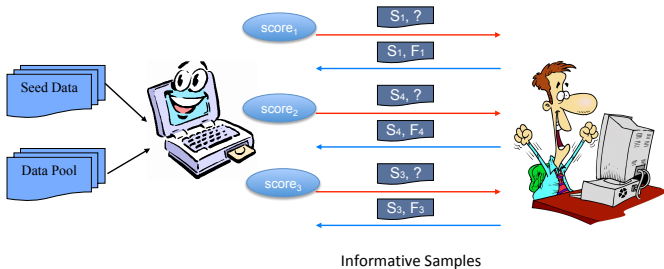
◀ نوعی یادگیری ماشینی کم نظارت

◀ هدف این نوع یادگیری استخراج نمونه های حاوی اطلاعات جدید است.



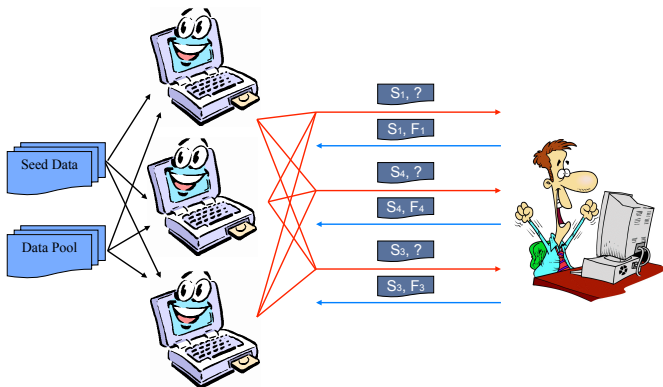
انواع یادگیری فعال (I)

نمونه گیری بدون قطعیت



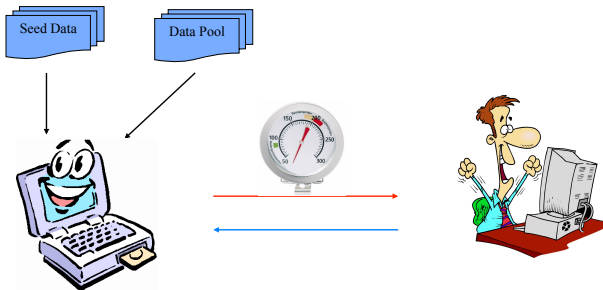
انواع یادگیری فعال (II)

پرسشگر گروهی



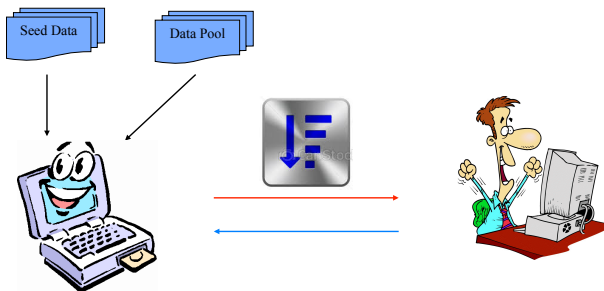
انواع نمونه گیری (I)

◀ نمونه گیری خطی



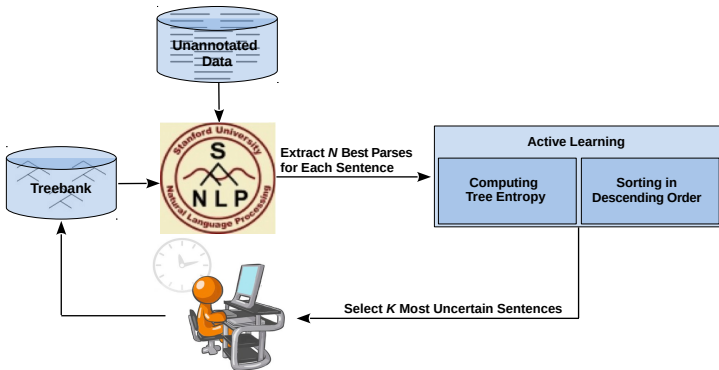
انواع نمونه‌گیری (II)

- ◀ نمونه‌گیری خطی
- ◀ نمونه‌گیری حجمی



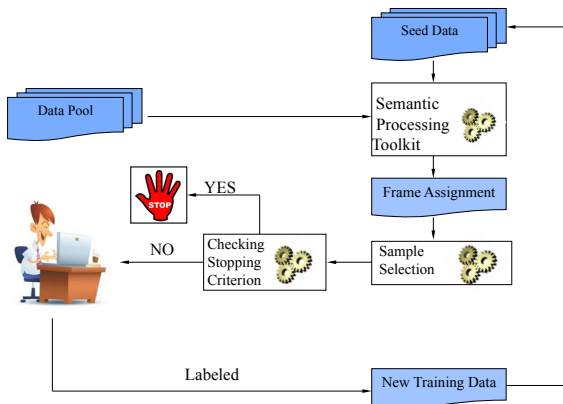
نمونه استفاده عملی از یادگیری فعال (I)

❖ قیومی (۲۰۱۴)



نمونه استفاده عملی از یادگیری فعال (II)

❖ قیومی (۲۰۰۹)



استفاده از یادگیری فعال (IV)

- ◀ نقطه قوت
- ◀ کاربرد ماشین برای نشانه‌گذاری داده

استفاده از یادگیری فعال (IV)

◀ نقطه قوت

◀ کاربرد ماشین برای نشانه‌گذاری داده

◀ یافتن نمونه‌های حاوی اطلاعات جدید و عدم نیاز به نشانه‌گذاری تمام داده

استفاده از یادگیری فعال (IV)

◀ نقطه قوت

- ◀ کاربرد ماشین برای نشانه‌گذاری داده
- ◀ یافتن نمونه‌های حاوی اطلاعات جدید و عدم نیاز به نشانه‌گذاری تمام داده
- ◀ جلوگیری از نشانه‌گذاری داده‌های با حشو اطلاعاتی

استفاده از یادگیری فعال (IV)

◀ نقطه قوت

- ◀ کاربرد ماشین برای نشانه‌گذاری داده
- ◀ یافتن نمونه‌های حاوی اطلاعات جدید و عدم نیاز به نشانه‌گذاری تمام داده
- ◀ جلوگیری از نشانه‌گذاری داده‌های با حشو اطلاعاتی
- ◀ مؤثر در کاهش نسبی تلاش انسان برای نشانه‌گذاری داده

استفاده از یادگیری فعال (IV)

◀ نقطه قوت

- ◀ کاربرد ماشین برای نشانه‌گذاری داده
- ◀ یافتن نمونه‌های حاوی اطلاعات جدید و عدم نیاز به نشانه‌گذاری تمام داده
- ◀ جلوگیری از نشانه‌گذاری داده‌های با حشو اطلاعاتی
- ◀ مؤثر در کاهش نسبی تلاش انسان برای نشانه‌گذاری داده
- ◀ مؤثر در کاهش نسبی زمان و هزینه برچسب‌گذاری

◀ نقطه ضعف

- ◀ نیاز به داده آموزش نشانه‌گذاری شده

استفاده از یادگیری فعال (IV)

◀ نقطه قوت

- ◀ کاربرد ماشین برای نشانه‌گذاری داده
- ◀ یافتن نمونه‌های حاوی اطلاعات جدید و عدم نیاز به نشانه‌گذاری تمام داده
- ◀ جلوگیری از نشانه‌گذاری داده‌های با حشو اطلاعاتی
- ◀ مؤثر در کاهش نسبی تلاش انسان برای نشانه‌گذاری داده
- ◀ مؤثر در کاهش نسبی زمان و هزینه برچسب‌گذاری

◀ نقطه ضعف

- ◀ نیاز به داده آموزش نشانه‌گذاری شده
- ◀ نیاز حداقلی به دخالت انسان در نشانه‌گذاری داده

استفاده از یادگیری فعال (IV)

◀ نقطه قوت

- ◀ کاربرد ماشین برای نشانه‌گذاری داده
- ◀ یافتن نمونه‌های حاوی اطلاعات جدید و عدم نیاز به نشانه‌گذاری تمام داده
- ◀ جلوگیری از نشانه‌گذاری داده‌های با حشو اطلاعاتی
- ◀ مؤثر در کاهش نسبی تلاش انسان برای نشانه‌گذاری داده
- ◀ مؤثر در کاهش نسبی زمان و هزینه برچسب‌گذاری

◀ نقطه ضعف

- ◀ نیاز به داده آموزش نشانه‌گذاری شده
- ◀ نیاز حداقلی به دخالت انسان در نشانه‌گذاری داده
- ◀ گمراه‌شدن مدل یادگیرنده به واسطه داده‌های جدید اضافه‌شده به داده آموزش که متفاوت از داده آموزش قبلی است.

فهرست موضوعات

۱ مقدمه

۲ پیکره زبانی و زبان‌شناسی پیکره‌ای

۳ نشانه‌گذاری داده

۴ چالش‌ها

۵ جمع‌بندی

چالش مربوط به داده

- ◀ تداخل حروف و علائم فارسی و عربی: ی/ی ؛ ک/ک ؛ حتماً/حتمناً ؛ حتی /
حتا ؛ بنا/بنا ؛ رئیس/رییس ؛ امریکائی/امریکایی/آمریکائی/آمریکایی

چالش مربوط به داده

- ◀ تداخل حروف و علائم فارسی و عربی: ی/ی ؛ ک/ک ؛ حتماً/حتمن ؛ حتی /
حتا ؛ بنا/بنا ؛ رئیس/رییس ؛ امریکائی/امریکایی/آمریکائی/آمریکایی
- ◀ اشکال در واحدسازی
- ◀ در غیر این صورت / درغیراین صورت / درغیراینصورت

چالش مربوط به داده

- ◀ تداخل حروف و علائم فارسی و عربی: ی/ی ؛ ک/ک ؛ حتماً/حتمناً ؛ حتی /
 حتا ؛ بنا/بنا ؛ رئیس/رییس ؛ امریکائی/امریکایی/آمریکائی/آمریکایی
- ◀ اشکال در واحدسازی
- ◀ در غیر این صورت / در غیر این صورت / در غیر این صورت
- ◀ ویابته‌راست (و یا بهتر است)، مادر (ما در)، راداری (را داری)، وبا (و با)

چالش مربوط به داده

- ◀ تداخل حروف و علائم فارسی و عربی: ی/ی ؛ ک/ك ؛ حتماً/حتمن ؛ حتی /
 حتا ؛ بنا/بنا ؛ رئیس/رییس ؛ امریکائی/امریکایی/آمریکائی/آمریکایی
- ◀ اشکال در واحدسازی
- ◀ در غیر این صورت / در غیر این صورت / در غیر این صورت
- ◀ ویابهِتراست (و یا بهتر است)، مادر (ما در)، راداری (را داری)، وبا (و با)
- ◀ کزو ؛ کزین ؛ زوی ؛ مرا

چالش مربوط به داده

- ◀ تداخل حروف و علائم فارسی و عربی: ی/ی ؛ ک/ك ؛ حتماً/حتمن ؛ حتی /
 حتا ؛ بنا/بنا ؛ رئیس/رییس ؛ امریکائی/امریکایی/آمریکائی/آمریکایی
- ◀ اشکال در واحدسازی
- ◀ در غیر این صورت / در غیر این صورت / در غیر این صورت
- ◀ ویابهتر است (و یا بهتر است)، مادر (ما در)، راداری (را داری)، وبا (و با)
- ◀ کزو ؛ کزین ؛ زوی ؛ مرا
- ◀ سبک: اگر / گر / اگه

چالش مربوط به داده

- ◀ تداخل حروف و علائم فارسی و عربی: ی/ی ؛ ک/ک ؛ حتماً/حتمن ؛ حتی /
 حتا ؛ بنا/بنا ؛ رئیس/رییس ؛ امریکائی/امریکایی/آمریکائی/آمریکایی
- ◀ اشکال در واحدسازی
- ◀ در غیر این صورت / در غیر این صورت / در غیر این صورت
- ◀ ویابته‌راست (و یا بهتر است)، مادر (ما در)، راداری (را داری)، وبا (و با)
- ◀ کزو ؛ کزین ؛ زوی ؛ مرا
- ◀ سبک: اگر / گر / اگه
- ◀ کاهش عمق خط و وجود مشکل هم‌نگارگی: شیر ؛ کند ؛ کرم ؛ مهر

چالش مربوط به داده

- ◀ تداخل حروف و علائم فارسی و عربی: ی/ی ؛ ک/ک ؛ حتماً/حتمن ؛ حتی /
 حتا ؛ بنا/بنا ؛ رئیس/رییس ؛ امریکائی/امریکایی/آمریکائی/آمریکایی
- ◀ اشکال در واحدسازی
- ◀ در غیر این صورت / در غیر این صورت / در غیر این صورت
- ◀ ویابته‌راست (و یا بهتر است)، مادر (ما در)، راداری (را داری)، وبا (و با)
- ◀ کزو ؛ کزین ؛ زوی ؛ مرا
- ◀ سبک: اگر / گر / اگه
- ◀ کاهش عمق خط و وجود مشکل هم‌نگارگی: شیر ؛ کند ؛ کرم ؛ مهر
- ◀ کسره اضافه: خانه/خانه‌ی/خانه

چالش مربوط به داده

- ◀ تداخل حروف و علائم فارسی و عربی: ی/ی ؛ ک/ك ؛ حتماً/حتمن ؛ حتی / حتا ؛ بنا/بنا ؛ رئیس/رییس ؛ امریکائی/امریکایی/آمریکائی/آمریکایی
- ◀ اشکال در واحدسازی
- ◀ در غیر این صورت / در غیر این صورت / در غیر این صورت
- ◀ ویابته‌راست (و یا بهتر است)، مادر (ما در)، راداری (را داری)، وبا (و با)
- ◀ کزو ؛ کزین ؛ زوی ؛ مرا
- ◀ سبک: اگر / گر / اگه
- ◀ کاهش عمق خط و وجود مشکل هم‌نگارگی: شیر ؛ کند ؛ کرم ؛ مهر
- ◀ کسره اضافه: خانه/خانه‌ی/خانه
- ◀ عدم رعایت استاندارد در فارسی‌نویسی واژه‌های لاتین: اینترمدیت، اینترمدیت، اینترمیدیت، اینترمیدیت

چالش مربوط به برچسب‌گذاری داده

◀ کاهش تطابق و انسجام در برچسب‌زنی و نیاز به بررسی سازگاری بین برچسب‌زنان

چالش مربوط به برچسب‌گذاری داده

- کاهش تطابق و انسجام در برچسب‌زنی و نیاز به بررسی سازگاری بین برچسب‌زنان
- محاسبه کاپا (کوهن، ۱۹۶۰)

$$k = \frac{p_o - p_e}{1 - p_e}$$

که p_o تعداد داده‌های داری توافق بین برچسب‌زنان و p_e تعداد داده‌های بدون توافق است.

چالش مربوط به برچسب‌گذاری داده

- کاهش تطابق و انسجام در برچسب‌زنی و نیاز به بررسی سازگاری بین برچسب‌زنان
- محاسبه کاپا (کوهن، ۱۹۶۰)

$$k = \frac{p_o - p_e}{1 - p_e}$$

که p_o تعداد داده‌های داری توافق بین برچسب‌زنان و p_e تعداد داده‌های بدون توافق است.

- تعدادی داده مشخص یکسان بین برچسب‌زنان وجود دارد تا امکان مقایسه داده وجود داشته باشد.

چالش مربوط به برچسب‌گذاری داده

- کاهش تطابق و انسجام در برچسب‌زنی و نیاز به بررسی سازگاری بین برچسب‌زنان
- محاسبه کاپا (کوهن، ۱۹۶۰)

$$k = \frac{p_o - p_e}{1 - p_e}$$

که p_o تعداد داده‌های داری توافق بین برچسب‌زنان و p_e تعداد داده‌های بدون توافق است.

- تعدادی داده مشخص یکسان بین برچسب‌زنان وجود دارد تا امکان مقایسه داده وجود داشته باشد.

- کاپا بین دو نفر محاسبه می‌شود. چنانچه تعداد برچسب‌زنان بیشتر از دو نفر باشد، دوبره‌دو افراد با هم مقایسه شده و از عدد کاپا میانگین گرفته می‌شود.

چالش مربوط به برچسب‌گذاری داده

- کاهش تطابق و انسجام در برچسب‌زنی و نیاز به بررسی سازگاری بین برچسب‌زنان
- محاسبه کاپا (کوهن، ۱۹۶۰)

$$k = \frac{p_o - p_e}{1 - p_e}$$

که p_o تعداد داده‌های داری توافق بین برچسب‌زنان و p_e تعداد داده‌های بدون توافق است.

- تعدادی داده مشخص یکسان بین برچسب‌زنان وجود دارد تا امکان مقایسه داده وجود داشته باشد.
- کاپا بین دو نفر محاسبه می‌شود. چنانچه تعداد برچسب‌زنان بیشتر از دو نفر باشد، دوبره‌دو افراد با هم مقایسه شده و از عدد کاپا میانگین گرفته می‌شود.
- $60\% < k \leq$ ضعیف

چالش مربوط به برچسب‌گذاری داده

- کاهش تطابق و انسجام در برچسب‌زنی و نیاز به بررسی سازگاری بین برچسب‌زنان
- محاسبه کاپا (کوهن، ۱۹۶۰)

$$k = \frac{p_o - p_e}{1 - p_e}$$

که p_o تعداد داده‌های داری توافق بین برچسب‌زنان و p_e تعداد داده‌های بدون توافق است.

- تعدادی داده مشخص یکسان بین برچسب‌زنان وجود دارد تا امکان مقایسه داده وجود داشته باشد.

- کاپا بین دو نفر محاسبه می‌شود. چنانچه تعداد برچسب‌زنان بیشتر از دو نفر باشد، دوبه‌دو افراد با هم مقایسه شده و از عدد کاپا میانگین گرفته می‌شود.

$60\% < k \leq$ ضعیف

$60\% < k \leq$ خوب

چالش مربوط به برچسب‌گذاری داده

- کاهش تطابق و انسجام در برچسب‌زنی و نیاز به بررسی سازگاری بین برچسب‌زنان
- محاسبه کاپا (کوهن، ۱۹۶۰)

$$k = \frac{p_o - p_e}{1 - p_e}$$

که p_o تعداد داده‌های داری توافق بین برچسب‌زنان و p_e تعداد داده‌های بدون توافق است.

- تعدادی داده مشخص یکسان بین برچسب‌زنان وجود دارد تا امکان مقایسه داده وجود داشته باشد.

- کاپا بین دو نفر محاسبه می‌شود. چنانچه تعداد برچسب‌زنان بیشتر از دو نفر باشد، دوبره‌دو افراد با هم مقایسه شده و از عدد کاپا میانگین گرفته می‌شود.

ضعیف $k < 60\%$

خوب $k > 60\%$

عالی $k > 80\%$

چالش مربوط به مدل یادگیرنده

◀ عدم وجود یادگیرنده ثابت

چالش مربوط به مدل یادگیرنده

- ◀ عدم وجود یادگیرنده ثابت
- ◀ نیاز به تهیه مدل یادگیرنده با توجه به نوع برچسب مورد نظر

چالش مربوط به مدل یادگیرنده

- ◀ عدم وجود یادگیرنده ثابت
- ◀ نیاز به تهیه مدل یادگیرنده با توجه به نوع برچسب مورد نظر
- ◀ نیاز به امکانات پردازشی رایانشی

چالش مربوط به مدل یادگیرنده

- ◀ عدم وجود یادگیرنده ثابت
- ◀ نیاز به تهیه مدل یادگیرنده با توجه به نوع برچسب مورد نظر
- ◀ نیاز به امکانات پردازشی رایانشی
- ◀ صرف زمان برای تعامل بین ماشین برای یافتن نمونه داده‌های دشوار و انسان برای برچسب‌گذاری این داده‌ها

فهرست موضوعات

مقدمه

۱

پیکره زبانی و زبان‌شناسی پیکره‌ای

۲

نشانه‌گذاری داده

۳

چالش‌ها

۴

جمع‌بندی

۵

جمع‌بندی

◀ معرفی پیکره و زبان‌شناسی پیکره‌ای

جمع‌بندی

- ◀ معرفی پیکره و زبان‌شناسی پیکره‌ای
- ◀ نحوه گردآوری داده

جمع‌بندی

- ◀ معرفی پیکره و زبان‌شناسی پیکره‌ای
- ◀ نحوه گردآوری داده
- ◀ نحوه نشانه‌گذاری داده به صورت دستی و ماشینی

جمع‌بندی

- ◀ معرفی پیکره و زبان‌شناسی پیکره‌ای
- ◀ نحوه گردآوری داده
- ◀ نحوه نشانه‌گذاری داده به صورت دستی و ماشینی
- ◀ بررسی نقاط ضعف و قوت ارائه نحوه نشانه‌گذاری داده: به نظر می‌رسد یادگیری فعال روش مناسبی برای نشانه‌گذاری داده در حداقل حجم و کاهش زمان و هزینه باشد.

جمع‌بندی

- ◀ معرفی پیکره و زبان‌شناسی پیکره‌ای
- ◀ نحوه گردآوری داده
- ◀ نحوه نشانه‌گذاری داده به صورت دستی و ماشینی
- ◀ بررسی نقاط ضعف و قوت ارائه نحوه نشانه‌گذاری داده: به نظر می‌رسد یادگیری فعال روش مناسبی برای نشانه‌گذاری داده در حداقل حجم و کاهش زمان و هزینه باشد.
- ◀ بررسی چالش‌های مربوط به داده، برچسب‌زنی و مدل یادگیرنده: به نظر می‌رسد علیرغم چالش‌ها، برچسب‌گذاری ماشینی بسیار بهتر از برچسب‌گذاری دستی باشد. این چالش‌ها در مقایسه با صرف زمان و هزینه برچسب‌گذاری دستی ناچیز است.

