

Morfologické značkování je jednou ze základních a zásadních úloh v oblasti zpracování přirozeného jazyka. Morfologické značkovače trénované metodami řízeného strojového učení fungují dobře pro jazyky, pro které existují velká ručně anotovaná data, např. angličtina, francouzština, portugalština, atp. Pro ostatní jazyky nelze metody řízeného strojového učení použít. V této práci trénujeme morfologický značkovač metodou neřízeného strojového učení na vícejazyčných paralelních datech, která jsou použita pro přenos morfologické informace z jednoho (zdrojového) jazyka do druhého (cílového). Naše metoda dosahuje výsledků srovnatelných se současnými nejlepšími metodami (porovnání provedeno na 8 jazycích), ale používá výrazně méně trénovacích dat a je jednodušší, což má za následek výrazně větší rychlost zpracování. V práci se dále zabýváme otázkou optimální volby zdrojového jazyka. Ukazuje se, že Angličtina je optimální jen výjimečně. Naše metoda umí predikovat optimální zdrojový jazyk jen na základě jednojazyčných rysů. Při použití rysů z paralelních dat se kvalita predikce zlepšuje. V práci dále ukazujeme, že úspěšnost značkování se zlepšuje v případě kombinace více zdrojových jazyků.