Směry výzkumu v ÚFAL MFF UK od roku 2005

Tento text je určen pro interní potřebu pracovníků a doktorandů ÚFAL
Autor: Jan Hajič
Původní zdroj: http://ufal.mff.cuni.cz/internal/docs/vyzkum-ufal.html

1. Preambule - co tento dokument je a co není neboli jak jej číst

V tomto textu se pokusíme naznačit výzkumné směrování ÚFALu od roku 2005. Jsou v něm stanoveny dlouhodobé cíle i konkrétní směry rozvoje v teoretické i aplikované oblasti a rámec, v jakém bude možno provádět i další výzkumné nebo vývojové aktivity (spolupráce s průmyslem, s jinými institucemi, se zahraničím).

Motivací k vytvoření tohoto dokumentu byla snaha dát každému současnému (a po jeho zveřejnění i potenciálním budoucím) pracovníkům nebo studentům ÚFALu informaci o tom, co Ústav "chce" dělat (na základě toho, která témata jsou, a snad i nějakou dobu budou, z hlediska rozvoje oboru ve světě aktuální), ve kterých směrech bude poskytovat peníze z "centrálních" zdrojů (např. z výzkumného záměru, případně z jiných, relativně "volných" prostředků), a tedy ve kterých směrech jsou a budou prioritně zajištěny peníze na další rozvoj. Měl by dát možnost (po ustálení současné situace a doplnění o "mapu" toho, co je a není pokryto současnými studenty a pracovníky), aby každý zájemce našel oblast, ve které by mohl buď sám, nebo (a to zejména) v týmu s daným zaměřením pracovat a podílet se jak na teoretické práci, tak na tvorbě algoritmů a nástrojů. Do "hlavních" směrů výzkumu se postupně jistě dostanou i další témata, neboť rozvoj oboru je stále bouřlivý a nelze detailně předvídat vývoj najednou na více než několik let. Ústav je přitom průběžně otevřen jakémukoli tématu z oboru počítačové lingvistiky nebo tématu s ní související za podmínek specifikovaných v tomto dokumentu. Priority výzkumu v rámci tohoto dokumentu (které zahrnuje mnohem více, než jsme schopni v současné době pokrýt finančně i personálně) bude určovat vedení Ústavu po konzultaci s akademickými a vědeckými pracovníky Ústavu s přihlédnutím k možnostem Ústavu.

Celý tento text byl psán z hlediska budoucí prosperity ÚFALu jako celku, jeho dobrého postavení v ČR i ve světě, a jeho dalšího rozvoje, nikoli z hlediska současného personálního obsazení Ústavu. Každý však má šanci se v seznamu uvedených konkrétních teoretických i praktických úloh "najít" (koneckonců tento seznam je mnohem větší, než náplň práce ÚFALu a CKL v současnosti). Celkově objem práce ovšem je a bude limitován množstvím "centrálních" financí a rozsahem probíhajících cílených projektů.

Tento dokument nemá sloužit jako návod k tomu, jak organizovat a podávat granty (vč. zahraničních) v budoucnu. K tomuto velmi důležitému tématu bude ale v krátké době vytvořen podrobně zvláštní dokument, který bude vycházet z tohoto textu a z obecně známých grantových možností a pravidel, a stanoví strategii a taktiku při získávání dalších prostředků.

Tento dokument se rovněž netýká výuky, byť ta je nedílnou a důležitou součástí náplně práce ÚFALu (i když jsme "Ústav" a ne "katedra", jsme součástí vysoké školy a nikoli čistě vědeckým ústavem jako jsou např. ústavy Akademie). Výuka přitom vychází z potřeb

výzkumu (v oblasti výchovy doktorandů) a z kombinace potřeb výzkumu a praxe (na bakalářském a magisterském stupni). Vzhledem ke komplexnosti problému výuky a výchovy vědeckých pracovníků bude k tomuto tématu vytvořen zvláštní dokument (rovněž v co nejbližší době).
Některé úlohy uvedené v tomto dokumentu jsou dnes již pokryty některými probíhajícími nebo začínajícími projekty, zejména tzv. Informační společnosti, ale i granty GAČR, GAUK a mezinárodní spolupráce. Takové projekty a granty budou pokračovat, a to i tehdy, pokud nejsou přímo stanoveny jako prioritní v tomto dokumentu. Problematika "přechodného období" roku 2005 (a výjimečně možná i roku 2006) bude řešena individuálně s vedoucími jednotlivých projektů.

2. Motivace

Dlouhodobá prosperita jakéhokoli vědeckého pracoviště je závislá na jeho výsledcích. Bez ohledu na momentální prosazování různých "objektivních" metod "scientometrie" (nejen v ČR) to především znamená publikovat, a to v uznávaných časopisech a sbornících (seznam je přiložen; budeme doufat, že i při (formálním) hodnocení Ústavu "zvnějšku" bude tento seznam relevantní). Publikace (a vytvořená data a softwarové nástroje) by pak měly přinést i dobré hodnocení v Ústavu vedených projektů a tedy ve spojení s vhodnou strategií podávání grantů a projektů i snazší přístup k finančním zdrojům v budoucnu.

Abychom mohli publikovat, musíme mít výsledky, které zajímají komunitu počítačové lingvistiky. Ať už se na tuto komunitu díváme jako na těch několik stovek recenzentů nabídnutých příspěvků, nebo jako na všechny, kteří v oboru pracují a čtou, případně používají naše data a nástroje, jde nakonec o nejobjektivnější posouzení naší vědecké a výzkumné práce.

Dobré publikace mohou vznikat dvojím způsobem: prací skutečných vědeckých individualit (v oborech jako je matematika a společenské vědy), nebo společnou prací v týmech (obvykle pod vedením takových individualit - a to zejména v přírodních vědách, jako je fyzika a chemie). Počítačová lingvistika je dnes ve své hlavní části vědou empirickou a experimentální, a tedy má z tohoto hlediska mnohem blíže k fyzice než ke společenským vědám, ke kterým je někdy zařazována: její úlohy jsou velmi složité, potřebuje někdy velmi drahé zdroje (např. anotované korpusy) a řada úloh musí nutně navazovat na jiné, "state-of-the-art" výsledky jiných úloh (např. syntaktická analýza navazuje na morfologickou), neboť řešit každou úlohu od prvopočátků je neúnosně neefektivní. Proto následující výzkumné směry byly stanoveny poměrně konkrétně, jako podklad pro vytvoření týmů, které budou pracovat na jednotlivých úlohách, a v jejich rámci budou řešit pro komunitu počítačové lingvistiky přínosné a zajímavé problémy. Předpokládáme přitom, že Ústav bude mít kolem 20-25 (relativně) stabilních členů (pracovníků pedagogických, vědeckých a odborných včetně doktorandů zaměstnaných na projektech). Výsledky jednotlivých úloh budou přitom moci přebírat - díky jejich modulárnosti - i další týmy nebo skupiny Ústavu, které se tak k těm zajímavým výsledkům dostanou dříve a bez nutnosti duplikovat práci již hotovou. Mělo by se tím zamezit i problémům, kdy se výsledků dosáhne v "nevhodnou dobu" (tj. např. příliš "předčasné") nebo takovým způsobem, že tyto výsledky nezapadají do souvisejících projektů (např. tehdy, jestliže jejich výstup postrádá to, co potřebuje navazující program).

Přesto, že následující návrh se dívá především do budoucna, bere v úvahu dosavadní expertízu Ústavu jako celku, zejména v oblasti teorie a budování komplexně anotovaných korpusů (to pochopitelně předpokládá, že bude-li Ústav dále pokračovat, bude v něm nadále většina

současných pracovníků a doktorandů pracovat). Řada věcí, které by "bylo hezké" v návrhu mít, v něm ovšem z podobných důvodů chybí: nebudeme-li mít "kritickou" úroveň expertízy (= příslušně hluboce vzdělaných lidí), nelze se "ve velkém" pouštět do práce v oblastech, kde takovou expertizu a zároveň vysokou úroveň výsledků prokazují jiní - mohlo by tomu tak ale být např. ve spolupráci s nimi.

3. Celkové zaměření

1. **Dlouhodobý cíl**
   Ústav formální a aplikované lingvistiky je zaměřen na počítačovou lingvistiku včetně spojení oboru rozpoznávání mluvené řeči a zpracování přrozeného jazyka. Dlouhodobým cílem Ústavu je - v souladu s cíli oboru - popsat chování jazykového systému (a konkrétně češtiny a angličtiny, případně dalších jazyků) tak, aby bylo bezprostředně možno vytvořit automatické nástroje (počítačové programy) pro automatickou analýzu mluveného či psaného sdělení, jejímž výsledkem bude formalizovaná struktura reflektující obsah tohoto sdělení, a obráceně, vytvořit vnější formu sdělení reprezentovaného takovou formální strukturou. Ústav přitom v případě mluveného jazyka přejímá výsledky v oblasti akustického zpracování signálu, případně jiných tzv. modalit (zpracování obrazu a jiných signálů z vnějšího světa) od partnerských pracovišť v ČR i zahraničí. Úkolem Ústavu je i vytvářet výzkumné aplikace, které demonstrují možnosti poskytované výše uvedenými nástroji pro potenciální aplikace využívající v jakékoli míře zpracování přirozeného jazyka.
2. **Střednědobý cíl**
   Hlavním výzkumným programem Ústavu v příštích 4-6 letech bude

   a. v teoretické oblasti: rozvoj explicitního popisu chování jazyka, mluveného i psaného, s přesahem do obsahu jazykových sdělení, integrované v rámci Pražského závislostního korpusu;
   b. v aplikované oblasti: vývoj samostatně použitelných, ale navazujících programových nástrojů "od povrchu k obsahu" (a zpět), integrovaní do systému strojového překladu.

   Výzkum v jiné oblasti a výzkum a vývoj pro průmysl bude možný při splnění podmínek uvedených níže v kap. 4 o zdrojích na výzkum.
3. **Metodologie**
   Pro vývoj jednotlivých nástrojů na zpracování přirozeného jazyka budou používány takové metody, které povedou k nejlepším výsledkům podle jim odpovídajících světově přijatých a z hlediska publikační činnosti standardních kritérií vyhodnocování. K účelu hodnocení podle těchto metrik budou vyvíjena nebo získávána potřebná testovací data. Bude podporováno i získávání a případně i tvorba vlastních dat pro účely trénování algoritmů pro strojové učení (do kterých zahrnujeme i algoritmy založené na pravděpodobnostních i jiných statistických modelech), pokud bude na základě dostupné literatury a předchozích experimentů existovat oprávněná a podložená domněnka, že se jedná o efektivní přístup k řešení daného problému. Data vytvořená vlastními silami před jejich publikací vždy interně projdou procesem zajištění vysoké kvality (formálně i obsahově), včetně kvantitativního vymezení hranic kvality manuální anotace (shoda anotátorů, konzistence "horizontální" i "vertikální" atd.).

Teoretický výzkum, nutně omezený vzhledem k použitelným prostředkům bude vždy směřovat k tomu, aby byly postupně vyvinuty dokonalejší nástroje a systémy (byť i v dlouhé časové perspektivě).

Bude kladen důraz na to, aby přejaté postupy (včetně využití dat vytvořených mimo ÚFAL) nebránily jak širokému výzkumnému, tak i případnému průmyslovému využití vytvořených nástrojů. Ústav bude podporovat šíření vytvořených dat a nástrojů takovou formou, která umožní jejich zpřístupnění co nejširšímu okruhu zájemců; v případě dat jde o nízkou cenu pro výzkumné využití (blízkou nákladům na vlastní šíření dat) a v případě nástrojů především o vhodnou politiku licencování s minimálními restrikcemi pro vědecké a výzkumné využití všemi zájemci.

4. **Zdroje na výzkum**
   Vedení Ústavu bude podporovat teoretický i aplikovaný výzkum v níže popsaných hlavních směrech tím, že jednak podpoří (i personálně) podání takových cílených projektů, které budou tyto směry pokrývat, a jednak bude poskytovat přímo finanční prostředky z "centrálních" zdrojů Ústavu (především jde o rozpočet ÚFALu a Výzkumný záměr IS).
   Ostatní výzkum a vývoj bude umožněn pod hlavičkou Ústavu za těchto podmínek:

   a. příslušnost k oboru počítačová lingvistika
   b. realistická možnost publikace výsledků na světové úrovni nebo velká prestiž projektu ve světě
   c. dostatečné finanční krytí (a to včetně infrastruktury; projekt nesmí vyžadovat dodatečné zdroje Ústavu - bylo by naopak vhodné (i když ne nezbytně), aby pokryl ve formě jisté "daně" např. i další výzkumnou činnost).

   Základní strategie získávání finančních zdrojů ze všech zdrojů na výzkum je popsána ve zvláštním dokumentu [finance-ufal.doc].

5. **Lidské zdroje**
   Výuka studentů a výchova vědeckých pracovníků je nedílnou a důležitou součástí práce ÚFALu. Jedná se o komplexní problém, který je popsán ve zvláštním dokumentu [vyuka-ufal.doc].

4. Teoretický výzkum

Teoretický výzkum slouží k přípravě "půdy" pro vývoj nástrojů, anotovaných dat a systémů. Veškeré úsilí tedy musí mít cíl, byť i dlouhodobě, pro který daný výzkum slouží. V tom je zahrnuta i případná anotace dat jako poslední předstupeň před vývojem nástrojů a systémů. Výzkum bude zahrnovat následující oblasti:

a. specifikace obsahové reprezentace popisu jazykových sdělení
b. pojmenované entity (zahrnuje číselné a jiné výrazy pro kvantitu)
c. lexikální význam a jeho rozlišování, terminologie, frazeologie, vztah ontologie a lexikální sémantiky (a to i v souvislosti s obsahovou reprezentací)
d. (ko)reference v jazyce (jako problém ležící mezi rovinou tektogramatickou a rovinou obsahovou), problematika sdílení znalosti mezi mluvčím a posluchačem, otázky dynamiky dialogu
e. specifické problémy přechodu od vnější formy k významu a obsahu pro mluvený jazyk a obráceně (včetně dopadu na jazykové modelování pro automatické rozpoznávání řeči a syntézu řeči)

f. specifikace tektogramatické roviny vč. výše uvedených dílčích problémů pro angličtinu, příp. další jazyky
g. metodika komplexního anotování jazykových (textových i mluvených) korpusů
h. formální popis jazyka a metody lingvistiky.

Do teoretické části výzkumu spadá i výzkum v oblasti slovníků v rámci jednotlivých výše uvedených bodů, avšak jejich vlastní tvorba bude probíhat podle potřeb vývoje nástrojů.

5. Systémy a nástroje, aplikovaný výzkum

Programové nástroje a systémy jsou v ÚFAL vyvíjeny tak, aby umožnily jak publikační činnost a srovnání s obdobnými špičkovými nástroji v zahraničí i ČR, tak snadný přechod k aplikacím.

Ústav si stanovil jako střednědobý cíl vytvořit nebo upravit sadu vzájemně navazujících a propojených nástrojů na analýzu a syntézu přirozeného jazyka ve formě programových modulů. Součástí střednědobého cíle je sjednocení rozhraní mezi moduly ve smyslu použitých externích datových struktur i předávání řízení a parametrů jednotlivým modulům, jako předpoklad snadné integrace do větších celků.

Jako příklad vzorového výzkumného systému, ve kterém bude integrovaná většina vyvinutých nástrojů, byl vybrán strojový překlad z angličtiny do češtiny a obráceně. Systém bude postaven modulárně, aby na sebe jednotlivé nástroje přímo navazovaly. Tyto nástroje budou přitom využitelné i samostatně. Všechny součásti systému budou budovány v rámci Ústavu, případně bude využito existujících zdrojů a nástrojů tak, aby byla umožněna výzkumná i průmyslová spolupráce (vč. zajištění práv pro šíření).

Všechny vytvořené nástroje bez výjimky budou vyhodnoceny samostatně podle všobecně uznávaných kvantitativních kritérii. Zároveň bude možné příspěvek jednotlivých modulů vyhodnocovat i v systému překladu jako celku.
Nástroje budou tedy splňovat tato kritéria:

- modularita (jasně definovaná funkčnost, implementace v jednotném prostředí ve formě knihoven / programových celků);
- návaznost na další nástroje (formát dat, způsob komunikace mezi moduly);
- kvantitativní vyhodnotitelnost objektivními (publikovatelnými) postupy.

Mezi základní systémy (které obvykle mají i několik různých podsystémů, které se dále mohou skládat z mnoha specializovaných nástrojů), zpracovávání na ÚFAL, tedy patří:

a. Morfologie - analýza i syntéza, tagging: postupné zdokonalování systému
b. Parsing (analytický, český a anglický)
   Cílem je mít vlastní parser, vč. analytických funkcí.
c. Tektogramatický parsing (tj. s výstupem na úrovni roviny tektogramatické, český a anglický)
   Tektogramatický parsing je možné řešit jednak jako "transformační", s využitím parsingu na úrovni analytické roviny, tak i bez ní (tj. se vstupem přímo z morfologické analýzy a/nebo taggingu).
d. Lexikální disambiguace, jména a jiné "named entities", čísla a obecně kvantity (čeština, angličtina)

e. Odkazování v textech (čeština, angličtina)
f. Analýza aktuálního členění (českého, anglického)
g. Transfer (vlastní strojový překlad): anglicko-český i česko-anglický
h. Generování (anglické, české)
i. Rozpoznávání a syntéza řeči
   Bude prováděno ve spolupráci s jinými pracovišti; nebude se řešit "signálová" část rozpoznávání/syntézy řeči.
j. Nástroje pro anotaci a vyhledávání

6. Průmyslové a zahraniční spolupráce

V této oblasti bude práce řízena víceméně "trhem", tj. jednání s firmami (s výjimkou firem, které by byly součástí CKAL) bude vedeno, pokud ony vyjádří zájem s námi spolupracovat. U zahraničních grantů budeme partnery v hlavních směrech výzkumu sami vyhledávat, a budeme uvažovat o spolupráci i tehdy, bude-li nám nabídnuta mimo hlavní směry našeho výzkumu (v EU je v tomto ohledu nevýhodná situace, neboť NLP není samostatnou prioritou). Nadále budeme (i přes absenci vhodného formálního rámce, jakým jsou v EU tzv. rámcové programy) spolupracovat s vedoucími pracovišti v USA. Současné společné projekty se zahraničím a současná spolupráce s průmyslem budou pokračovat.

Přestože Ústav tedy nebude aktivně vyvíjet koncové aplikace, bude aktivně propagovat výsledky své vědecké práce i mezi potenciálními uživateli v průmyslové sféře. Bude k tomu využívat existujících osobních kontaktů a bude tyto kontakty dále získávat. Ústav bude rovněž propagovat svou práci v médiích v maximálně možné míře.