Ablation Study

This section report a detailed analysis of entity recognition capabilities of machine-learning classifiers involved in all of the transfer learning experiments. First column of every tables indicates the input space representation that can be one between **Contextualized Hidden States (hidden), Attention Scores (attention)** and **Source Probability Distribution (probs)** or a concatenation of these representations represented by the operator +.

Conll -> Ontonotes

Model	pa	svm	sgd	per	mlp	bma
attention	0.7311	0.7622	0.741	0.646	0.7634	0.7690
hidden	0.753	0.6701	0.7728	0.7459	0.8015	0.807
probs	0.447	0.5189	0.5276	0.411	0.5461	0.5506
Att+hidden	0.7953	0.6658	0.7950	0.7713	0.8118	0.8261
Attention+probs	0.7122	0.7622	0.7351	0.5300	0.7531	0.7675
Probs+hidden	0.64	0.5893	0.7574	0.789	0.8164	0.7761
Att+probs+hidden	0.8073	0.6715	0.7957	0.7574	0.8035	**0.8263**

Table 1: Predictive performance from ConLL03 to Ontonotes5.0 with 5000 training instances. Scores represent f1-score of machine ing models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	bma
attention	0.6904	0.6775	0.6886	0.6751	0.7494	0.7417
hidden	0.7295	0.7117	0.7342	0.7196	0.7859	0.7286
probs	0.4776	0.5180	0.5329	0.5299	0.5464	0.5319
Att+hidden	0.7660	0.6974	0.7763	0.7419	0.7972	0.7540
Att+probs	0.6745	0.7344	0.7258	0.6643	0.7432	0.7413
Probs+hidden	0.7633	0.7171	0.7725	0.7521	0.7988	0.7194
Att+probs+hidden	0.7761	0.6991	0.7834	0.7675	0.7959	0.7457

Table 2: Predictive performance from ConLL03 to Ontonotes5.0 with 2500 training instances. Scores represent f1-score of machine ing models trained on specific vector representation

Model	pa	svm	sgd	per	mlp	bma
attention	0.6984	0.7174	0.6984	0.5738	0.7318	0.7385
hidden	0.6775	0.7060	0.175	0.5190	0.7218	0.7218
probs	0.4971	0.5182	0.5292	0.1026	0.5457	0.5272
Att+hidden	0.7094	0.696	0.6786	0.6061	0.7107	0.743
Att+probs	0.639	0.7179	0.711	0.642	0.740	0.7087
Probs+hidden	0.6775	0.7107	0.1777	0.6153	0.7091	0.7156
Att + probs + hidden	0.6726	0.6947	0.7035	0.6624	0.6764	0.7411

Table 3: Predictive performance from ConLL03 to Ontonotes5.0 with 1200 training instances. Scores represent f1-score of machine ing models trained on specific vector representation

Ontonotes -> conll

Model	Bert + pa	Bert + svm	Bert + sgd	Bert + per	Bert + mlp	Bert + bma
attention	0.8472	0.8652	0.8753	0.8693	0.8680	0.8792
hidden	0.7062	0.6053	0.2433	0.7790	0.8132	0.6954
probs	0.7699	0.7793	0.775	0.7002	0.7823	0.7776
Att+hidden	0.8171	0.6195	0.7914	0.8031	0.7736	0.7746
Att+probs	0.8529	0.8651	0.8536	0.8360	0.8755	0.8714
Probs+hidden	0.7494	0.5607	0.2083	0.7314	0.7963	0.7054
Att + probs + hidden	0.7990	0.6187	0.7632	0.7964	0.8229	0.7735

Table 4: Predictive performance from Ontonotes5.0 to ConLL03 with 5000 training instances. Scores represent f1-score of machine ing models trained on specific vector representation.

Model	Bert + pa	Bert + svm	Bert + sgd	Bert + per	Bert + mlp	Bert + bma
attention	0.8316	0.8525	0.8411	0.8337	0.8558	0.8659
hidden	0.7612	0.6571	0.2401	0.7276	0.8152	0.7244
probs	0.7496	0.7785	0.7742	0.5718	0.7819	0.7749
Att+hidden	0.8195	0.6613	0.8088	0.8199	0.8351	0.7667
Att+probs	0.8496	0.8532	0.8521	0.8279	0.8638	0.8650
Probs+hidden	0.7977	0.5819	0.2249	0.6921	0.8194	0.7127
Att + probs + hidden	0.8102	0.6625	0.8085	0.7662	0.8230	0.7721

Table 5: Predictive performance from Ontonotes5.0 to ConLL03 with 2500 training instances. Scores represent f1-score of machine ing models trained on specific vector representation

Model	Bert + pa	Bert + svm	Bert + sgd	Bert + per	Bert + mlp	Bert + bma
attention	0.8559	0.8558	0.8377	0.8531	0.8606	0.8649
hidden	0.7498	0.6725	0.2568	0.7259	0.7726	0.7286
probs	0.7663	0.7778	0.7773	0.6148	0.7838	0.7767
Att+hidden	0.8238	0.8068	0.8115	0.8473	0.8563	0.8195
Att+probs	0.8448	0.8555	0.8342	0.8389	0.8566	0.8629
Probs+hidden	0.7819	0.7075	0.2680	0.7651	0.7824	0.7562
Att + probs + hidden	0.8274	0.8068	0.8318	0.8018	0.8552	0.8163

Table 6: Predictive performance from Ontonotes5.0 to ConLL03 with 1200 training instances. Scores represent f1-score of machine ing models trained on specific vector representation.

Multinerd -> CoNLL

Model	pa	svm	sgd	per	mlp	bma
attention	0.8255	0.8343	0.8119	0.8143	0.8386	0.8434
hidden	0.4840	0.4148	0.3914	0.3636	0.6285	0.5132
probs	0.7561	0.7432	0.7596	0.7424	0.7605	0.7678
Att+hidden	0.6568	0.4706	0.5063	0.6729	0.7065	0.6919
Att+probs	0.8300	0.8345	0.7807	0.8106	0.8343	0.8371
Probs+hidden	0.4465	0.4089	0.4643	0.4905	0.5820	0.5549
Att + probs + hidden	0.7420	0.4737	0.6828	0.7310	0.7856	0.7194

Table 7: Predictive performance from Multinerd to ConLL03 with 1200 training instances. Scores represent f1-score of machine ing models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	bma
attention	0.8066	0.8259	0.8306	0.7830	0.8428	0.8406
hidden	0.4621	0.2020	0.4932	0.5553	0.4287	0.5681
probs	0.7305	0.7509	0.7729	0.6086	0.7581	0.7704
Att+hidden	0.5716	0.3443	0.5965	0.6624	0.5727	0.6437
Att+probs	0.8132	0.8255	0.8192	0.8093	0.8469	0.8434
Probs+hidden	0.4845	0.1999	0.4305	0.5307	0.4641	0.5538
Att + probs + hidden	0.6804	0.3454	0.6558	0.6956	0.7269	0.5586

Table 8: Predictive performance from Multinerd to ConLL03 with 2500 training instances. Scores represent f1-score of machine ing models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	bma
attention	0.8355	0.8429	0.8353	0.8032	0.8384	0.8520
hidden	0.5814	0.4059	0.5391	0.5512	0.4663	0.6355
probs	0.7278	0.7482	0.7702	0.6663	0.7739	0.7714
Att+hidden	0.6614	0.211	0.6485	0.6882	0.6831	0.6923
Att+probs	0.8228	0.8423	0.8409	0.8151	0.8414	0.8560
Probs+hidden	0.6206	0.3499	0.6002	0.5471	0.5200	0.6427
Att + probs + hidden	0.6894	0.1976	0.6294	0.6687	0.5625	0.6699

Table 9: Predictive performance from Multinerd to ConLL03 with 5000 training instances. Scores represent f1-score of machine ing models trained on specific vector representation.

CoNLL -> Multinerd

Model	pa	svm	sgd	per	mlp	bma
attention	0.8941	0.9025	0.9040	0.8693	0.9020	0.9032
hidden	0.7668	0.8605	0.1943	0.8378	0.8504	0.7789
probs	0.8637	0.8739	0.8499	0.8432	0.8638	0.8641
Att+hidden	0.8353	0.8739	0.8712	0.8169	0.8742	0.8806
Att+probs	0.8931	0.9024	0.8959	0.8777	0.9063	0.9088
Probs+hidden	0.8314	-	0.1806	0.8291	0.8618	0.7817
Att + probs + hidden	0.8423	0.8739	0.8244	0.8277	0.8695	0.8850

Table 10: Predictive performance from ConLL03 to Multinerd with 5000 training instances. Scores represent f1-score of machine ing models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	bma
attention	0.8855	0.8940	0.8929	0.8846	0.8954	0.8987
hidden	0.8084	0.8708	0.1805	0.8077	0.8273	0.8738
probs	0.8353	0.859	0.8550	0.8408	0.8609	0.8646
Att+hidden	0.8754	0.8671	0.8707	0.8479	0.8453	0.8783
Att+probs	0.8908	0.8944	0.8870	0.8724	0.8992	0.9005
Probs+hidden	0.8362	0.8723	0.1711	0.7689	0.8510	0.8731
Att + probs + hidden	0.8896	0.8878	0.8286	0.8392	0.8795	0.8725

Table 11: Predictive performance from ConLL03 to Multinerd with 2500 training instances. Scores represent f1-score of machine ing models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	bma
attention	0.8765	0.8899	0.8636	0.8701	0.8882	0.8915
hidden	0.8035	0.8708	0.1809	0.7995	0.7887	0.8660
probs	0.8536	0.8547	0.8537	0.8156	0.8548	0.8578
Att+hidden	0.7587	0.8390	0.8275	0.8176	0.8461	0.8762
Att+probs	0.8635	0.8899	0.8864	0.8870	0.8929	0.8932
Probs+hidden	0.7587	0.8716	0.8275	0.8176	0.8461	0.8642

Att + probs + hidden	0.8775	0.8420	0.8354	0.8081	0.8607	0.8759

Table 12: Predictive performance from ConLL03 to Multinerd with 1200 training instances. Scores represent f1-score of machine ing models trained on specific vector representation.

Ontonotes5.0 -> Multinerd

Model	pa	svm	sgd	per	mlp	bma
attention	0.8934	0.8976	0.8956	0.8602	0.9038	0.9009
hidden	0.6483	0.5329	0.5001	0.6235	0.7382	0.7547
probs	0.7484	0.7827	0.7698	0.7648	0.7839	0.7849
Att+hidden	0.7953	0.4987	0.8008	0.8063	0.7463	0.8197
Att+probs	0.8847	0.8974	0.8999	0.8616	0.8994	0.8197
Probs+hidden	0.6250	0.5401	0.4535	0.6286	0.7503	0.7684
Att + probs + hidden	0.7797	0.4693	0.7942	0.7818	0.7567	0.8052

Table 13: Predictive performance from Ontonotes5.0 to Multinerd with 5000 training instances. Scores represent f1-score of machine ing models trained on specific vector representation

Model	pa	svm	sgd	per	mlp	bma
attention	0.8709	0.8880	0.8736	0.8589	0.8914	0.8861
hidden	0.6789	0.6009	0.5843	0.6479	0.7022	0.6246
probs	0.7556	0.7812	0.7716	0.7142	0.7829	0.7819
Att+hidden	0.8329	0.4001	0.8091	0.7626	0.8107	0.4584

Att+probs	0.8789	0.8879	0.8773	0.8830	0.8853	0.8859
Probs+hidden	0.6636	0.5920	0.5752	0.6366	0.6556	0.6151
Att + probs + hidden	0.8186	0.4194	0.8068	0.7896	0.8250	0.4787

Table 14: Predictive performance from Ontonotes5.0 to Multinerd with 2500 training instances. Scores represent f1-score of machine ing models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	bma
attention	0.8643	0.8794	0.8644	0.8488	0.8825	0.8777
hidden	0.6338	0.3941	0.4387	0.5902	0.7035	0.4701
probs	0.7280	0.7831	0.7696	0.5590	0.7838	0.7831
Att+hidden	0.8254	0.7465	0.7839	0.8027	0.8024	0.7593
Att+probs	0.8406	0.8791	0.8646	0.8713	0.8814	0.8778
Probs+hidden	0.6061	0.4093	0.4288	0.5548	0.6662	0.4815
Att + probs + hidden	0.8124	0.7523	0.7622	0.8048	0.8340	0.7652

Table 15: Predictive performance from Ontonotes5.0 to Multinerd with 1200 training instances. Scores represent f1-score of machine ing models trained on specific vector representation.

Multinerd ontonotes

Model	pa	svm	sgd	per	mlp	bma
attention	0.6625	0.7432	0.7271	0.6469	0.7591	0.7619
hidden	0.4319	0.3050	0.4749	0.3996	0.4663	0.5299
probs	0.4302	0.4981	0.5008	0.4074	0.5314	0.5079
Att+hidden	0.4775	0.1834	0.4675	0.5140	0.6178	0.5982
Att+probs	0.6538	0.7435	0.7228	0.6713	0.7538	0.7465

Probs+hidden	0.3271	0.3103	0.4590	0.4132	0.5550	0.5427
Att + probs + hidden	0.5320	0.1619	0.4869	0.5846	0.6446	0.6372

Table 16: Predictive performance from Multinerd to Ontonotes5.0 with 5000 training instances. Scores represent f1-score of machine ing models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	bma
attention	0.6974	0.7168	0.7122	0.6816	0.7324	0.7452
hidden	0.4602	0.2595	0.4656	0.4547	0.4781	0.5234
probs	0.4539	0.4978	0.5061	0.4650	0.5269	0.5097
Att+hidden	0.5281	0.3199	0.5131	0.5993	0.5948	0.6253
Att+probs	0.6840	0.7175	0.7042	0.6561	0.7186	0.7433
Probs+hidden	0.4062	0.2568	0.4549	0.4309	0.4270	0.5859
Att + probs + hidden	0.5511	0.3315	0.4886	0.6151	0.6423	0.6345

Table 17: Predictive performance from Multinerd to Ontonotes5.0 with 2500 training instances. Scores represent f1-score of machine ing models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	bma
attention	0.6069	0.6907	0.6779	0.6314	0.7313	0.7228
hidden	0.4192	0.1896	0.4579	0.4086	0.4833	0.5220
probs	0.1036	0.4964	0.4985	0.1060	0.5273	0.5053
Att+hidden	0.5574	0.3905	0.5301	0.5701	0.6030	0.6281
Att+probs	0.6817	0.6927	0.7007	0.6432	0.7271	0.7250

Probs+hidden	0.3854	0.1851	0.5106	0.3975	0.4323	0.535
Att + probs + hidden	0.4875	0.4045	0.5315	0.5252	0.6218	0.6249

Table 18: Predictive performance from Multinerd to Ontonotes5.0 with 1200 training instances. Scores represent f1-score of machine ing models trained on specific vector representation.

Contribution of different input space representations

Regarding the contribution of each component of the vector space, we can say that Attention Scores, derived from the multi-head attention weights from the final layer of a pre-trained language model, is the representation that gives more contribution to the goal of learn the objective function (eq 1 of the paper). Only the transfer from CoNLL03 to Ontonotes 5.0 show that **Contextualized Hidden States** gives more contribution.

However, Contextualized Hidden States, that is denoted by the hidden state representation from the final layer of a pre-trained language model, give less but not less important contributions to the f1-score. Source Probability Distributions is the feature less contributive for the task due to the complexity of distinguishing between tokens with the same probability distribution but labeled with a tag not present in the source domain.

Concatenating features input space not always gave benefits to the task, tables show same trend, concatenation of Attention Scores and Source Probability Distributions can achieve more robust performance comparated with other concatenation experiments, but compare to the solely attention scores input space representation predictive performance drops. Additionally concatenation of features increase training and inference time due to the huge dimensionality of vectors that represent single tokens.

Contribution of machine-learning classifiers

Tables show a trend where Multi-Layer Perceptron and SVM achieve best predictive performance into respect other machine-learning classifiers. MLP and SVM gives more contribution also to the Bayesian Model Averaging (BMA) Ensemble. Is not always true that BMA, that in this ablation study takes all classifiers trained on the same input space, gain predictive performance into respect the classifiers that compose it.

In some case predictions errors with an high probability score inducing ensemble model to vote the wrong entity.